課題研究報告書

文書の特徴語抽出に関する技術の調査と実験

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

井内 寛

2011 年 12 月

課題研究報告書

文書の特徴語抽出に関する技術の調査と実験

指導教官　島津 明 教授

審査委員主査　島津 明 教授
審査委員　東条 敏 教授
審査委員　白井 清昭 准教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

0710951 井内 寛

提出年月: 2011 年 11 月

概 要

現代において世の中には大量の情報が溢れかえっており、やり取りされる文書の情報量は、近年増加の一途をたどっており、効率的な情報収集に努めなければいけない。

このような効率的な情報の収集を行おうとしたとき、検索という方法により、必要な情報を探し出すことになる。検索によって必要な情報へと素早くたどりつくためには、事前に整理された情報の手引きの存在や、必要な情報が既にまとめられている必要がある。だが、それらの作成にはコストがかかり、また、真に必要な情報であるかどうかを事前に知ることは困難である。

機械的に特徴語を抽出することができれば、事前での情報の収集に大きく役立つことになり、機械的な特徴語抽出の手法に関して調査・考察することは重要である。

また、このような大量の文書があふれかえっている分野として、法律制定の分野が挙げられる。法律文の電子データ・検索システムが公開されているが、法律単位でのインデックスが不十分であり、現状では、検索に対して十分とはいえない。また、法律文の特徴語の辞書化やタグ付けなどの資源の整備も進んでいない。

そこで、各特徴語抽出における手法を用いて法律文に適用し、問題点を洗い出し評価することは大いに重要であると考えられる。よって、本研究において特徴語抽出の先行研究の調査を行い、その手法を実際の法律文に適用し、考察することとする。

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　研究の目的

現代において、世の中には大量の情報が溢れかえっている。人口の増加と通信手段の発達に伴い、一日あたりにやり取りされる文書の情報量は、近年増加の一途をたどっており効率的な情報収集に努めなければいけない。

ユーザが情報を収集を行おうとしたとき、大量の文書から必要かどうかを判断することは現実的には困難である。したがって、その時ユーザは検索という方法を取り、必要な情報に関係する情報を探し出すことになる。検索により得られた文書は、ランク付けもされているため、ユーザは順に目を通していき、目的の情報を手に入れることができる。

一般的に検索はクエリ（問い合わせ）によって行われる。すなわち、ユーザは自らが調べたいことを理解し、調べたい事柄を検索のクエリとして認識していなければならない。検索を行うことは、情報を選別することであり、必要な情報へと素早くたどりつくことに役立つ。

しかし欠点として、ユーザが必要とする情報にたどり着くためのクエリをあらかじめ知っていなければいけない点が挙げられる。このような場合、整理された情報の手引きの存在や、必要な情報が既にまとめられてあれば何も問題はなく、ユーザは必要とする情報であると感じる部分に目を通せばよい。だが、それらの作成にはコストがかかり、また、真に必要な情報であるかどうかを事前に知ることは困難である。

機械的に特徴語を抽出することができれば、事前での情報の収集に大きく役立つことになり、機械的な特徴語抽出の手法に関して調査・考察することは重要である。そこで本研究では、文書の特徴語抽出に関する技術の調査と主な手法による実験を行い、問題点を洗い出し評価することを研究の目的とする。

本課題研究報告書は以下の構成で成る。まず2章で、特徴語抽出に関する技術の関連研究について説明する。次に3章で、主な手法の有効性を確認するための実験とその際に利用したデータについて述べる。最後に4章で、本課題研究報告書の結論と今後の課題を述べる。

1.2 研究の背景・重要性

特徴語抽出の技術がよく用いられる分野として、web 上のコンテンツなどがある。例えば livedoor blog [1] などのブログエントリーやニュースサイト cnet japan [2] などのニュース記事のように、web 上のコンテンツは「タグ」を付与し管理されることが多い。

文書の話題に合ったタグ付けを行えば、それを確認するだけで、対象となる文書の大まかな内容の把握が可能になり、また希望する情報に辿り着くのが容易となる。また、類似した内容に関する文書同士には、同じタグを付与することで、文書をカテゴリ分けすることができる。

しかし、理解の乏しい話題に関する文書であったり、特に明確なテーマがない文書など、タグを付与することを考えることが難しい場合などがある。また、たとえタグを考えたとしても、より適切なタグの候補が存在する可能性も考えられる。そのような場合、文書から特徴的な語を自動で抽出する特徴語抽出に関する技術が重要となってくる。

一般的に、文書から特徴語を抽出するには、文書中の単語に対して重要度の計算を行い特徴語を決定することになる。その特徴語抽出の方法には、多様な用途があり、また、重要度の計算方法においても、これまでの先行研究によって様々な手法が提案されている。

特徴語抽出の手法に関しては TF-IDF[1] が一般的に用いられている。この TF-IDF は、該当文書内に出現する語の頻度の情報をもとに重要度を決定するという特徴を持つ。

しかし、日本語などの言語では主題となる表現が頻繁に省略され、重要語とすべき語が頻出するとは限らないため、文書中の出現頻度をもとに重要度を求める TF-IDF が、必ずしも適切な重要度の値を出力するとは限らない。

また、杉浦 [2] は、見出し語となるような特徴的な語には、「主語や目的語になる」、「資料に出現する」、といった共通の性質があると考え、これらの傾向を特徴ベクトルといった異なる観点からなる特徴に対して定量化を行っている。そして、これらのベクトルの要素をもとに Support Vector Machine（SVM）[3] と呼ばれる識別器を用いた分類を行い特徴語の抽出を行っている。

著者の主張を表わす語をぬき出すことのできる特徴語抽出の手法として KeyGraph[4] がある。KeyGraph は、文書は著者独自の考えを主張するために書かれるという仮説をもとにしている。これは、文書を建物に例え、文書形成の準備あるいは前提となる基礎概念となる語の集合を土台、そして土台に強い力で支えられて文書を統合する語を屋根、土台と屋根を結ぶ強い力が集まった語を柱とした 3 点の概念を頼りに文書の特徴語の抽出を試みている。これは、頻出語の共起グラフを土台とし、この土台と共起する確率の高い語の集合を屋根としている。そして土台の語と屋根の語の関連度が高い語の集合を柱と捉え、これらを特徴語として抽出する。

また他にも、特徴語の抽出に語の共起関係を利用する研究は多く行われている。その中で、χ^2 検定の χ^2 値を用いた研究がある [5]。松尾らの研究では、文書中において重要な意味を持つ語は、共起する語に何らかの偏りがあると考えた。そこで、頻出語の出現割合

[1] http://blog.livedoor.com/
[2] http://japan.cnet.com/

と、頻出語との共起割合の間にどのくらいの偏りがあるかを χ^2 検定により調べた。χ^2 検定では、統計量 χ^2 を求めることにより、2 つの分布のずれを知ることができる。χ^2 値が大きければ、2 つの分布のずれが大きく、特徴語であると言える。松尾らの研究では、単純な出現割合ではなく、文の長さを考慮に入れて χ^2 値を求めている。

専門分野コーパスからの専門用語の抽出法として、湯本らは、単名詞 N に連接する単名詞の頻度の統計量を利用する N のスコア付けを一般的に表わす枠組みを提案している [6]。これらスコア付け方法を複合名詞のスコア付けに拡張し、比較として、既存の C-value[7] を修正した MC-value について述べている。これらのスコア付け法を NTCIR-1 TMREC タスクのテストコレクションに適用して結果を評価し、より包括的に（1,500〜10,000 語）専門語を抽出したいのなら、MC-value のほうが優れた結果を示すが、正解語を含む長めの語でよいのであれば、提案手法は大部分をカバーすることができ、スコア上位の候補においては提案手法の性能が優れてることを示した。

あらかじめ指定された情報を文書中から抽出することを目的とする情報抽出に関する特徴語抽出に、固有表現抽出というタスクがある。固有表現抽出は、文書に対し文書中の固有表現部分を抜き出し、抜き出した部分があらかじめ指定されたどの種別の固有表現であるかを分類するタスクである。

固有表現抽出は、情報抽出などの要素技術として、その重要性が指摘されており、入力文を適当な解析単位（トークン）に分割し、その単位に基づき固有表現部分をまとめあげるという手法が一般的である。トークンの単位として、単語や文字が考えられる。Asahara ら [8] は、文字を用いた手法が単語を用いた手法よりも高い抽出精度が得られることを示したが、この手法では、該当する文字の 2 文字程度の品詞情報のみを利用するため、固有表現の構成単語数が増加するにつれ、正確に固有表現を抽出することが難しくなる問題があった。

中野 [9] は、固有表現抽出の手法として、解析単位を文字単位だけでなく文節区切りまでも行い、文節内の情報を固有表現抽出のための素性として利用した。CRL 固有表現データ [10] を用いた評価実験の結果、F 値約 0.89 という結果を示し、提案手法の有効性を確認している。

1.3 研究の内容

このような大量の文書があふれかえっている分野として、法律制定の分野が挙げられる。我々の社会の構造や手続きは各種の法令によって規定されており、情報システムを規定する一種の仕様と見ることができる。また、社会の変化に対応して法令の制定・変更作業が頻繁で多大なコストが掛かっており、この作業の一部を計算機に行わせることは重要である。これは自然言語処理の様々な技術を用いることで支援することができる。

たとえば以下のような法律文があったとする。

> 障害基礎年金は、疾病にかかり、又は負傷し、かつ、その疾病又は負傷及びこれらに起因する疾病 (以下「傷病」という。) について初めて医師又は歯科医師の診療を受けた日 (以下「初診日」という。) において次の各号のいずれかに該当した者が、当該初診日から起算して一年六月を経過した日 (その期間内にその傷病が治つた場合においては、その治つた日 (その症状が固定し治療の効果が期待できない状態に至つた日を含む。) とし、以下「障害認定日」という。) において、その傷病により次項に規定する障害等級に該当する程度の障害の状態にあるときに、その者に支給する。ただし、当該傷病に係る初診日の前日において、当該初診日の属する月の前々月までに被保険者期間があり、かつ、当該被保険者期間に係る保険料納付済期間と保険料免除期間とを合算した期間が当該被保険者期間の三分の二に満たないときは、この限りでない。

この法律文が何を意味しているのか、一見しただけで把握することは難しい。しかし、この法律文中には「障害」、「疾病」、「傷病」、「障害認定日」などの、特徴的な語が含まれている。これらの語を見ることにより、この法律文が「障害」に関係していたり、「障害の支給要件」という法律についての規定をしていることが類推できる。このように特徴的な語を抽出し、それらをまとめた辞書を作成することができれば、ユーザがこの辞書を調べることで、必要な情報にたどり着くことができると考えられる。

また、このようにして得られた法律文の特徴語は、冒頭で述べたように同じ特徴語のタグを付与することで、法律文を分類することができる。上記の法律文で抽出した「障害」、「疾病」、「傷病」、「障害認定日」などの特徴語に関して、別の法律文で同じ特徴語が抽出された場合、同類の法律文として分類することができる。こうした類似法律文は、法令の変更の波及を調査する場合など、変更した法律文からの変更の波及がある可能性のある関連法律文候補として取得することが期待できる。

このように、特徴語を抽出することができれば、情報の収集に大きく役立ち、機械的な特徴語抽出の手法に関して調査・考察することは重要である。

現在、法律制定分野においては、法令データ提供システム「イーカブ」[3] などにおいて法律文の電子データ・検索システムが公開されているが、法律単位でのインデックスが不十分であり、現状では、検索に対して十分とはいえない。また、依然上記で述べたような

[3]イーカブ http://law.e-gov.go.jp/cgi-bin/idxsearch.cgi

法律文の特徴語の辞書化やタグ付けなどの資源の整備が進んでいない。そこで、様々な提案されている特徴語抽出における手法を用いて法律文に適用し、問題点を洗い出し評価することは、大いに重要であると考えられる。よって、本研究において特徴語抽出の先行研究を調査し、その手法を実際の法律文に適用し、考察することを目的とする。

第2章　特徴語抽出に関する方法

2.1　特徴語の定義

特徴語とひと言にいっても、その目的によって分類することができる。

まず、文書における情報検索のため文書を特徴付けることを目的とし、特徴語抽出を行うことが挙げられる。これは、索引付け（Indexing）と呼ばれる。索引付けにおいては、その文書を特徴付けるという性質である特定性と、文書をもれなく抽出するという性質である網羅性の関係が重要になってくる。

これらの性質は、情報検索における精度と再現率に関係してくる。特定性を高くするには、その文書に出現するが、他の文書には出現しないような語を抽出すればよいが、文書にあまりに特化した語だけを抽出すれば、検索クエリでその語が用いられる可能性も低くなってしまい、その文書が検索されにくくなってしまう。

また逆に、一般によく用いられる語を抽出すれば多くの文書の索引語となる可能性は高くなるが、検索クエリでこのような語を用いると多くの文書が検索されることになり、検索されたすべての文書が必ずしも必要としている文書であるとは限らなくなってしまう。このように情報検索における特徴語抽出の特定性と網羅性はトレードオフの関係といえ、これらのバランスをどのように取るかは、重要な研究課題だといえる。

情報検索の目的が、ユーザの要求に適合する文書を見つけだすことであるのに対して、あらかじめ指定された情報を文書中から抽出することを目的とする、情報抽出に関する特徴語抽出がある。これは、情報検索の検索クエリに比べ、どのような情報を抽出するかを詳細に指定する必要がある。

例えば、固有名や時間・数などの固有表現や照応関係の同定などが挙げられる。また、これらの情報抽出の手法として、人名や時間表現などそれぞれの固有表現に対して、それらの語の文字列のパターンを多数の正規表現などを使って抽出する、ヒューリスティックな手法や、予め固有表現のタグが付与されたタグ付きコーパスを、テストコレクションとして機械学習による統計的な手法で抽出を行う手法がある。

また、専門分野のコーパスから専門用語を自動的に抽出することを目的とした、特徴語抽出がある。専門用語の多くは複合語、とりわけ複合名詞であることが多く、名詞（単名詞と複合名詞）を対象として抽出を行う。複合名詞の専門用語は少数の基本的かつこれ以上分割不可能な名詞の組み合わせで形成されていることが多く、複合名詞とその要素である単名詞の関係性に着目することが重要となる。

これらの特徴語の抽出における手法は、すでに多くの先行研究がなされている。情報

検索を目的とした特徴語抽出の手法としては、TF-IDF[1] や杉浦の SVM を利用する手法 [2] や大澤らの KeyGraph[4]、松尾らの χ^2 値を用いる手法 [5] などが提案されている。また、情報抽出を目的とした手法では、中野の文節情報を利用した固有表現抽出の手法 [9] が提案されており、専門用語の抽出を目的とした手法では、湯本らによる出現頻度と連接頻度に基づく手法 [6] が提案されている。次節にて、これらの特徴語抽出の手法を紹介していく。

2.2 特徴語抽出に関する関連研究

2.2.1 TF-IDF による手法

一般に特徴語の索引付けには、TF-IDF が用いられことが多い。索引付けの主な役割は、文書中からその文書を特徴づける索引語を抽出することであるが、抽出した索引語がその文書の内容に、どれだけ関係しているかを重要度として定量化することが索引語の重み付けの尺度となる。そして索引語の重みを考えてみた場合、まず文書中における語の頻度が挙げられる。語の頻度に基づく重み付けの背景には、

> 何度も繰り返し言及される概念は重要な概念である [11]

という仮説がある。

しかし、高頻度の語に高い重み付け仮定してみると、一般的によく使われる語が重要ということになってしまい、文書を特徴付ける上ではあまり役に立たない。また、文書が長くなると平均的に語の出現頻度も高くなり、同じ索引語でも長い文書に現れる語の方が重みが大きくなってしまう問題がある。これは文書内の頻度は考慮しているが、文書集合全体の索引語の分布について考慮されていないためである。

ある索引語が、どの程度対象文書に特徴的に出現するかという特定性を考えた場合、他の文書中の索引語の分布も考慮する必要がある。このような特定性を表すための尺度として IDF（inverse document frequency）[12] がよく知られている。IDF はある索引語が全文書中に対してどれ位の文書に出現するかを表す尺度で、式 2.1 で定義される。

$$IDF(w) = \log\left(\frac{N}{DF(w)}\right) + 1 \tag{2.1}$$

ここで、N は文書集合中の全文書数であり、$DF(w)$ は語 w が出現する文書数になる。IDF はある語 t が少数の文書にしか出現しない場合に大きくなり、どの文書にもまんべんなく出現すると最小の値をとる。N と $DF(w)$ の比の対数をとるのは文書集合の大きさに対して IDF の値の変化量を抑えるためである。このように IDF は索引語の特定性を表現することができ、特定の少数の文書にしか出現しない語を捉える尺度となる。そして、上記で述べた語の頻度 $TF(w)$ と $IDF(w)$ を単独で用いるよりも、2 つの尺度を組み合わせて索引語の重みを計算することが考えられる。

具体的には式 2.2 のとおり、頻度 $TF(w)$ と $IDF(w)$ の積を用いる。

$$TF\text{-}IDF(w) = TF(w) \times IDF(w) \tag{2.2}$$

しかし、文書から得られる特徴語の数が文書量に大きく依存したり、この式で計算される重要性は、文書頻度が高ければ高いほど IDF が小さくなり、結果として式 2.2 も低い値となって、重要性は低いとされてしまう。つまり、文書グループ内でのみ文書頻度が高い単語においては、文書グループを考慮した重要性と TF-IDF が示す重要性とが無関係であったり、逆の傾向を示すという問題が発生する。したがって、いくつかの語が連接する

複合語などの出現頻度が低い語は、本来重要な語であるにもかかわらず下位にあることが多い。

この問題に対し相澤 [23] の研究では、複数の観点に基づくランキング手法の算出法を提案している。相澤の手法では、複合語を適切に評価するために構成する語の関係や、語の追加によって新たな語となるかなどを評価するための尺度の定義を行っている。それは、「結合度」、「前接度」、「出現度」、「後接度」、「文脈度」、「重要度」と定義される尺度となり、これらの尺度に基づきランキングを行う。この手法を用いることで、多くの語が連結した複合語でも、重要な語であれば上位にランキングされるようになる。語を適切にランキングすることは、特徴語抽出に有効であると考えられる。例えば、特徴語候補をランキングし、あるランク以下の語を足切りすることができれば、より精度の高い特徴語抽出ができると考えられる。

2.2.2 SVM を利用する手法

杉浦 [2] は、所属する研究室で行われるゼミを記録した議事録集合を対象として特徴語の抽出を行っている。また、特徴語に共通する性質を手掛かりとして、特徴語は話題の中心となることが多く、発言の中では主語や目的語として使われ、それは前後の語の品詞として助詞が出現しやすいと言い換えることができ、また、見出し語には特徴語が多く出現するという仮定を行っている。そして、このような特徴語の性質を用いて、まず特徴語になりそうな語を特徴語候補として選び出し、その候補を前述の特徴語としての性質を持っている語、持っていない語の二つに分類し、特徴語の性質を持っている語を特徴語として抽出を行っている。

このような性質による分類を行うために、各性質を定量化し特徴ベクトルとして考え、前後の品詞による要素と、スライド中への出現数という要素の 2 つの観点からなる情報の特徴ベクトルを設計している。このように、全く別の観点からなる情報を組み合わせることで、特徴語としての性質がより顕著に表れるのではないかと期待される。その特徴ベクトルを用いた分類には、Support Vector Machine（SVM）[3] と呼ばれる識別器を用いている。SVM は高次元の特徴ベクトルを分類するのに適した識別器である。この SVM により、特徴語候補を特徴語と非特徴語に分類することで特徴語抽出を行っている。

この提案されている特徴語抽出の詳細な手順を以下に説明する。まず最初に、文書中から語を取り出すために形態素解析を行い、文書の形態素の集合と品詞情報を取得する。形態素解析により得られた形態素は、語を構成する最小単位であり、これらの形態素の連結により、新たな語となることがある。そこで、特徴語候補の元になる語の集合を作るために、形態素の連結処理を行う。連結して新たな語となるかどうかは、その形態素の品詞により判別する。また、形態素解析は常に正確な分割を行うとは限らない。「特徴語」などの複合名詞などの場合は「特徴」と「語」に分割されてしまう。そこで、名詞が連接している場合は連結処理を行う。これにより、複合語などの二つ以上の形態素からなる語を特徴語の候補として選ぶことができるようになる。

連結処理により、特徴語候補のもととなる集合を取得し、この集合の中からいくつかの制約によるフィルタリングを行い、特徴語候補を抽出する。このフィルタリングのルールとして動詞や形容詞は、一般的な動作や形容を表すための語であるという考えより、動詞や形容詞などは特徴語として扱わないこととし、名詞を特徴語の対象とする。また、名詞の中でもいくつかの小分類が存在し、その中でも「接尾」、「非自立」、「代名詞」といった品詞がある。これらの形態素が語頭に来る語は、語としての条件を満たせていないとの考えから、先頭の形態素として「接尾」、「非自立」、「代名詞」が来ている語は、特徴語候補からの除外を行う。最後に、ある語の一部となる語の除去を行う。これは、特定の語と連結することが多い語は、単独で特徴語となることはないとの考えから、ある語 s の出現数の 90% 以上の割合で w を含む語 x の出現があった場合、w は x の一部であるとする。この場合 w は特徴語候補から除外する。

次に、特徴ベクトルの設定を行う。特徴語の言語的性質を考えた場合、特徴語となるような語は、その文書中での話題となり、主語や目的語になることが多い。主語や目的語となる場合は、その前後に助詞の出現が多いことから、特徴ベクトルの要素として前後の品詞の出現割合を利用する。しかし、前後の品詞による要素だけでは、文書に出現する回数に大きく依存してしまい、文書量に依存しない分類を行うために、文書から得られる見出しなどの特徴を特徴ベクトルの要素として取り入れる。このようにして各特徴語候補に対して特徴ベクトルの設定を行う。

また、提案手法では、特徴語抽出の対象となる議事録集合をいくつかのプロジェクトに分けている。各プロジェクトは全くの無関係ではないが、扱う対象が大きく異なり、それぞれのプロジェクトごとに特徴語となる語が違うため、特徴語候補には各プロジェクトごとに別々の特徴ベクトルの設定を行い、そのプロジェクトごとに分類を行い特徴語抽出を行う。

次に、これらの特徴ベクトルを利用し、特徴語候補から特徴語を見つけ出すためにSVMによる分類を行う。SVM での分類はカーネル関数を指定する必要がある。カーネル関数の取り方により分類に差が出るが、一般に最適なカーネル関数を求めることは難しい。提案手法では、最も適用範囲が広いと言われるガウシアン型カーネル [3] を用い分類を行っている。

また、提案手法では、特徴語候補を特徴語クラスと非特徴語クラスに分類している。この分類評価のために TF-IDF による特徴語抽出との比較を行っている。TF-IDF による特徴語抽出は、各特徴語候補の TF-IDF 値を求め、降順にソートし、その上位何語かを特徴語とするものである。

すべての特徴語候補、特徴語クラスに分類された語、非特徴語クラスに分類された語、それぞれの TF-IDF で抽出された特徴語の上位 200 語中に実際の特徴語が含まれている割合の比較を行う。これは、もし提案手法による特徴語抽出が有効であるなら、1. 特徴語クラス、2. すべての特徴語候補、3. 非特徴語クラス、の順に実際の特徴語を含む割合が大きいとの考えからである。

また、実際の特徴語かどうかの判断は、学習データとの一貫性を保持するため、特徴

語抽出の対象となる議事録集合を、対象プロジェクトのメンバーにより、特徴語であるかどうかの判断を行っている。その結果、TF-IDF による特徴語抽出よりも高い精度で特徴語を抽出することを示した。各プロジェクトごとの特徴語候補の語数、特徴語クラスの語数、非特徴語クラスの語数と、それぞれ上位 200 語が含む特徴語の割合を表 2.1 に示す。[1]

表 2.1: TF-IDF との比較による評価

プロジェクト名	A	B	C	D	E	F	G	H
AT	346	71	35.5%	1523	35	17.5%	45	22.5%
DM	570	88	44.0%	1889	38	19.0%	50	25.0%
VA	424	73	36.5%	1514	28	14.0%	39	19.5%

2.2.3 KeyGraph による手法

大澤らが提案している KeyGraph[4] は、文書中に出現する単語の出現頻度と共起関係からグラフ構造を作成し、そのグラフより文書の主張点を把握し特徴語を抽出する手法であり、文書は著者独自の考えを主張するために書かれるという仮説をもとにしている。

これは、文書を建物に例え、文書形成の準備あるいは前提となる基礎概念となる語の集合を土台、そして土台に強い力で支えられて文書を統合する語を屋根、土台と屋根を結ぶ強い力が集まった語を柱、とした 3 点の概念を頼りに、文書の特徴語の抽出を試みている。

これは、文書中で繰り返し出現する頻度の高い語は、その文書が書かれる上で前提とされる文書全体の、内容展開の基本となる概念である土台となることが多い。そして文書中では、この土台に基づいて文書に筋道が与えられる。この筋道に支えられているのが、文書中で筆者が最も伝えたい主張となり、この主張が文書の特徴語と成り得るという考えを基礎としている。以下に KeyGraph の詳細な手順を説明する。

KeyGraph では、

1. 土台の形成
2. 主張の抽出

[1] A: 特徴語クラスに分類された語数
B: 特徴語クラスの上位 200 語中の特徴語の語数
C: 特徴語クラスの上位 200 語中の特徴語の割合
D: 非特徴語クラスに分類された語数
E: 非特徴語クラスの上位 200 語中の特徴語の語数
F: 非特徴語クラスの上位 200 語中の特徴語の割合
G: 全特徴語候補の上位 200 語中の特徴語の語数
H: 全特徴語候補の上位 200 語中の特徴語の割合

の２つのフェーズからなる。以下、特徴語抽出の対象となる文書を D とし、文書 D における単語の共起関係を表すグラフを G と定義する。

最初に文書 D における出現頻度の上位語集合 $HighFreq$ を取り出す。この $HighFreq$ の要素をグラフ G のノード群とする。出現頻度の情報のみで $HighFreq$ とすると明らかに特徴語の候補として相応しくない語を含む可能性があり、これらをストップワードとして対象文書の語の集合から削除する。

次に、英語の場合はステミングを行い、日本語の場合は見出し語化を行う。また、単語の並びの組み合わせから熟語となる候補を生成し、熟語中に含まれる熟語の候補を出現回数で熟語候補から捨て、こうして残った熟語候補を熟語とする。

次に、 $HighFreq$ 中で文書 D における共起度の高い語の対を、それぞれ枝 (リンク) で結ぶ。ここで語の対の共起度 $co(w_i, w_j)$ は、式 2.3 のように文 s における語の出現回数の積の総和で定義される。[2]

$$co(w1, w2) = \sum_{s \in D} |w1|_s |w2|_s \tag{2.3}$$

共起度の範囲を文単位にすることにより、文の倒置や疑問文による語順の変化や複数文にまたがり共起するのを抑制し、精度向上を狙っている。また、上記の共起度を測る尺度の他に、相互情報量 [13] によって 2 語間の独立性を測る方法が挙げられるが、独立に出現する場合に比べて、近くに現れる回数自体が多い語の対を選ぶ方が適切と捉えたため、共起度に $co(w_i, w_j)$ を採用している。

次に、グラフ G 中の対になるノード w_i, w_j を結ぶ枝に対して、この枝を切り離したとしても、他の枝を遷移して w_iから w_j へと到達できる枝は、そのまま残し、到達できない枝は切断する。これは、極大連結部分グラフのみを残すことになる。こうして共起度 $co(w_i, w_j)$ の高い語の対の枝から得られたグラフ G 中の極大連結部分グラフを、文章形成の基礎概念として土台 PG とする。また、KeyGraph では、枝で結ばれている部分グラフだけでなく、独立したノードも 1 つの土台 PG として扱う。

文書から取り出したい特徴語は、土台に基づいて文書に筋道が与えられ、この筋道に支えられる語であるとの考えより、語 w が土台たちに支えられる力を $key(w)$ と定義する。[3]

$key(w)$ は、最初に語 w と土台 PG との共起度 $co2(w, PG)$ を計算する。[4]

$co2(w, PG)$ のスコアが高い語は、出現頻度の上位語集合 $HighFreq$ の語だけとは限らない。文書 D において $HighFreq$ とならなかった、出現頻度の低い語が土台 PG と強く共起する場合は、このような語をグラフ G に加える。

---

[2] $|x|_s$ は文 s における要素 x の出現回数で、 x が語の場合に $|x|_s$ は文 s 中の語 x の出現回数になる。

[3] $key(w)$ の対象となる語 w は、文書 D における特徴語となり得る全ての語である。すなわち文書 D よりストップワードに含まれる語を除いた全ての語となる。

[4] $co2(w, PG)$ は、語 w と土台 PG が含まれる文の数となる。
土台 PG を構成する語が複数存在し、1 文中に複数出現する場合は 1 カウントとする。

以上の手順で、語 w とグラフ G に含まれるすべての土台 PG との $co2(w, PG)$ を計算し、そのスコアの和を $key(w)$ とする。こうして得られた $key(w)$ の高いいくつかの語を特徴語として抽出する。

大澤らは、KeyGraph の性能の評価実験を行っている。これは、KeyGraph によって得られる特徴語が文書の主張と成り得るかどうかを、様々な学術論文について KeyGraph での特徴語抽出結果を、各論文の著者に対して質問を行い、評価の回答を得ている。[5] そして、大澤らは得た評価データにより、式 2.4、式 2.5 の指標で KeyGraph の性能の評価を行い、TF-IDF による手法と同等の精度であることを示している。

1. 抽出された特徴語の十分さ

$$suff = \frac{|A \cap K|}{|A|} \tag{2.4}$$

2. 抽出された特徴語の必要性

$$ness = \frac{|A \cap K|}{|K|} \tag{2.5}$$

ここで、 A, K はそれぞれ以下の集合となる。

A : 著者の主張を表す特徴語の集合

K : KeyGraph によって得られた特徴語の集合

抽出された特徴語の学術論文の著者による評価を表 2.2 に示す。

表 2.2: 抽出された特徴語の学術論文の著者による評価

	TF-IDF	KeyGraph
$suff$	65 / 88	76 / 88
$ness$	159 / 239	274 / 310

また、KeyGraph では、出現頻度の高い語のみを特徴語として抽出するだけでなく、出現頻度の低い語であっても、文書の主張となる語を特徴語として抽出できることを示した。

[5]実際には多数の著者からの回答を得ることは難しく、23 人の著者からの回答の評価となる。

2.2.4 χ^2 値を用いる手法

対象とする文書だけの情報から、語の共起をもとに統計的な指標を用い特徴語を抽出する手法として、松尾ら [5] は χ^2 値を用いる手法を考案した。

元々、大量のコーパスを背景とした情報検索を目的とするインデキシングなどの特徴語抽出の手法では、TF-IDF をはじめ、様々な手法が用いられているが、近年大量の電子文書が蓄えられるにしたがって、その文書の内容を大まかに把握するという目的での特徴語抽出も重要になっている。

また、特に Web ページにおいてこのような電子文書が蓄えられているが、Web 上の電子文書はその多様性により、適切なコーパスの収集コストが高く、全ての状況において用意できるものではない。このような背景の中、文書単独での特徴語抽出の手法が重要となってくる。

松尾らの手法は、対象とする文書の頻出語を取り出し、その頻出語と各語の共起頻度を求め、共起頻度がどのくらい偏っているかを、その語が重要語であるかどうかの指標として用いることによって、単一文書だけから比較的高い精度で特徴語の抽出が可能になることが大きな特徴となる。以下に χ^2 値を用いる手法を説明する。

松尾らは語が共起することの定義を、文書中に出現する単語は、文ごとに句点やピリオドによって区切られており、同文中に出現する 2 つの語は 1 回共起していると考える。また、ひとつの文書が与えられたとき、語の出現頻度から頻出語を取り出すことができ、この頻出語の集合を G とする。そして、語の共起の頻度を集計することにより、文書中に出現する語の共起行列を作ることができる。この共起行列は、文書中に出現する語の数を N とすると $N \times N$ の対称行列であるが、ここでは頻出語上位 M 語に対応する列だけを抜きだし、 $N \times M$ 行列としている。

語 w が頻出語 $g \in G$ と全く独立に生起すると仮定するなら、語 w と語 $g \in G$ が共起する確率は、頻出語 $g \in G$ の出現確率に従うことになる。また、語 w と頻出語 $g \in G$ の間に意味的なつながりがあれば、この共起する確率は偏ることになる。したがって、ある語 w の頻出語 $g \in G$ に対する共起確率が、頻出語単独での出現確率から、どのくらい偏りがあるかを測れば、その語の重要度を表す指標となるという考えである。

このような、統計的に有意なずれの評価を行うために、分布の偏りを検定する方法として χ^2 検定 [15] がよく用いられる。松尾らは、共起する確率の分布の偏りの程度を示す指標として、この χ^2 値を利用した。この統計量 χ^2 値は以下の式 2.6 で与えられる。[6]

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g} \tag{2.6}$$

この χ^2 値を指標として、文書自身の全体的な傾向から大きく逸脱する特徴を持つ語であるかどうかを判断し、特徴語として取り出すことになる。

---

[6] 頻出語単独での生起確率を理論確率 $p_g(g \in G)$ とし、語 w と頻出語群 G の共起の総数を n_w 、語 w と語 $g \in G$ の共起頻度を $freq(w,g)$ とする。

$n_w p_g$ は、語 w と語 g の共起する期待頻度を表す。

さらに松尾らは、文書中の文の長さは様々であり、長い文に出現する語は他の語と共起しやすく、短い文に出現する語は他の語と共起しにくい傾向がある。そして、共起の範囲を一文としているので、文の長さが長ければそれだけ他の語と共起する確率は増えることになり、逆に、短い文にも関わらず共起しているときには、その関係はより強いと考える。また、頻出語中の特定の一語 $g \in G$ とだけ共起する語は、χ^2 値は高くなるが、重要な語であるというより、語 g に付随する語である場合がほとんどであると考えた。これらを考慮するために、式 2.7 にて統計量 χ^2 値の変更を行っている。[7]

$$\hat{\chi^2}(w) = \chi^2(w) - \max_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g} \tag{2.7}$$

松尾らは、被験者への質問形式で評価実験を行った。評価実験の対象となる文書は松尾らの論文自体である。この抽出された特徴語の、χ^2 値を計算した結果を表 2.3 に示す。[8]

表 2.3: 論文に対しての χ^2 値上位の語

順位	χ^2 値	頻度	ラベル
1	126.1	147	語
2	81.2	14	キーワード+抽出
3	68.1	12	χ^2+値
4	45.6	20	確率
5	42.7	5	頻出+語
6	40.7	5	文書+キーワード+抽出
7	38.7	29	出現
8	35.0	5	分野
9	34.6	5	低い
10	34.3	17	語+共起

また、この手法の特徴語抽出の精度を評価するための評価実験として、人工知能の分野の 7 著者 20 論文に対して行い、「TF」、「TF-IDF」、「KeyGraph」との比較を行っている。これは、各手法で特徴語 15 語を抽出し、各手法から得られた特徴語の上位語をシャッフルし、各著者に、重要な概念を表すと思う語の評価の質問を行い、精度の評価を行っている。

各手法による出力語中で、特徴語であると判定された割合が precision である。また、提示した全ての語のうち、論文中で不可欠な概念を表す語を 5 つ以上選び A、B、C、D、E

[7] p_g を (g が出現する文数) / (G 中の語が出現する延べ文数) ではなく、(g が出現する文の語数の合計) / (文書全体の語数の合計) とする。
n_w を語 w が出現する文の語数の合計とする。

[8] 表中の「+」はフレーズを表している。

と印をつけ、それと同義の語にも同じ印をつける指示を行い、5つ以上の概念のうち各手法で提示した語にいくつ含まれているかで coverage の測定を行っている。

また、TF や TF-IDF は文書中でよく出てくる語の重みを大きくする傾向があり、抽出される特徴語は当たり前の語が多くなる。それに対し、この手法では出現頻度が少なくても重要な語を取り出すことが可能である。それを定量化したものが frequency index となり、これは抽出された語の出現頻度の平均を表す。この各手法の比較結果を表 2.4 に示す。

表 2.4: 各手法に対しての precision と coverage と frequency index

	TF	KeyGraph	χ^2 値による手法	TF-IDF
precision	0.53	0.42	0.51	0.55
coverage	0.48	0.44	0.61	0.61
frequency index	28.6	17.3	11.5	18.1

この結果、大量の文書の統計データを必要とする TF-IDF に匹敵する性能が得られた。また、TF や TF-IDF での手法は文書中でよく出てくる語の重みを大きくし、出てくる語は当たり前の語が多い中、それらに対し、松尾らの手法では出現頻度が少ない場合でも重要な語を取り出すことが可能であることを示した。

2.2.5 出現頻度と連接頻度に基づく手法

専門分野のコーパスから、専門用語を自動的に抽出することを目的とした特徴語抽出がある。専門用語の多くは名詞を組み合わせた複合名詞であることが多く、単名詞と複合名詞を対象として抽出を行う。複合名詞の専門用語は少数の基本的かつ、これ以上分割不可能な名詞の組み合わせで形成されていることが多く、複合名詞とその要素である単名詞の関係性に着目することが重要となる。

また、専門用語のもうひとつの重要な性質として、ある言語的単位の持つ分野固有の概念への関連性の強さであるターム性があげられる。ターム性とは、ある言語的単位の持つ分野固有の概念への関連性の強さと定義され、ターム性は専門文書を書いた専門家の概念に直結していると考えられる。このターム性を定量化するにあたり、ある単名詞が対象分野の重要な概念を表しているのなら、新たな専門用語を作り出す書き手はその単名詞を単独で使うのみならず、新たな概念を表す表現として、その単名詞を含む複合名詞を作りだすことが考えられる。このことから、複合名詞と単名詞の関係性を考慮することが重要になってくる。

中川ら [15] は、この関係性について、単名詞の前あるいは後に連接して複合名詞を形成する単名詞の種類数を使った、複合名詞の重要度スコア付けを提案している。しかしながら、この手法では、ある単名詞に連接する単名詞の頻度情報を考慮しておらず、ある程度コーパスが、出現する複合名詞の種類数が収束する程度に大きくなれば、頻度に影響され

ないので、スコアリングは一定の値になってしまう。そこで、湯本ら [6] は、単名詞に連接する単名詞の頻度の統計量を利用するスコア付け方法を提案した。

具体的には、特定のコーパスを想定し、単名詞 N のバイグラムをとったときに単名詞の左方にくる単名詞の種類の異なり数を n とし、単名詞の右方にくる単名詞の種類の異なり数を m とする。中川らは、この指標を単名詞 N のスコアとしていたが、湯本らは、この連接単名詞の異なり数の他に、それぞれ単名詞 N の左方、右方に連接して複合名詞を形成する全単名詞の頻度情報を取り入れ定義した。そして、複合名詞のスコアは複合名詞の長さに依存しないという考えの下、定義した各単名詞の左右のスコアの平均をとり、以下の式 2.8 にて複合名詞のスコア付けを行っている。[9]

$$LR(CN) = (\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1))^{\frac{1}{2L}} \tag{2.8}$$

そして、用語候補である単名詞あるいは複合名詞が単独で出現した頻度を考慮し、先の $LR(CN)$ と組み合わせ、式 2.9 でスコア付けを定義した。[10]

$$FLR(CN) = f(CN) \times LR(CN) \tag{2.9}$$

これらのスコア付け法を、既存の C-value 法 [16] を修正した MC-value 法と、連接種類数を用いた LR 法 (以下「連接種類 LR 法」) と、連接頻度を用いた LR 法 (以下「連接頻度 LR 法」) と、単名詞、複合名詞の単独での出現頻度をスコアとする語頻度法との比較を、NTCIR-1 タスクのテストコレクション[11] に適用して評価を行っている。ここで、C-value 法は、式 2.10 で定義される。[12]

$$C\text{-}value(CN) = (length(CN)-1) \times \left(n(CN) - \frac{t(CN)}{c(CN)}\right) \tag{2.10}$$

---

[9] 単名詞 N の左方のスコア関数を $FL(N)$ 、右方のスコア関数を $FR(N)$ 、単名詞 $N_1, N_2, .., N_L$ がこの順で連接した複合名詞が CN となる。

[10] $f(CN)$ は候補語 CN が単独で出現した頻度となる。

[11] NTCRIR-1 の TMREC タスクで利用されたテストコレクションである。

1999 年に行われた NTCIR-1 のタスクのひとつであった TMREC では、日本語のコーパスを配布して用語抽出を行う課題が行われた。

主催者側が人手で準備した用語に対して参加システムが抽出した用語の一致する度合いを評価した。

ただし、これらは何らかの客観的定量的基準に基づいて人手で選択されたものではなく、抽出者の直観によるものである。

[12] ここで、

CN : 複合名詞

$length(CN)$: CN の長さ (構成単名詞数)

$n(CN)$:コーパスにおける CN の出現回数

$t(CN)$: CN を含むより長い複合名詞の出現回数

$c(CN)$: CN を含むより長い複合名詞の異なり数

である。

しかし、式 2.10 では、 $length(CN) = 1$ である場合、すなわち CN が単名詞である場合に、C-value のスコアが 0 となり、適切なスコアリングができないため、湯本らは、C-value 法の定義を式 2.11 のように変更し、評価を行っている。

$$MC\text{-}value(CN) = length(CN) \times \left(n(CN) - \frac{t(CN)}{c(CN)} \right) \tag{2.11}$$

また、評価方法として、NTCRIR-1 の TMREC テストコレクションとして供給された正解用語[13] と比較し、抽出正解用語数、適合率、再現率、F 値を計算し、評価を行っている。これらは次式で定義される。[14]

$$\begin{aligned}
\text{抽出正解用語数}\,(PN) &= \text{上位 } PN \text{ 候補中の正解用語数} \\
\text{適合率}\,(PN) &= \frac{\text{抽出正解用語数}\,(PN)}{PN} \\
\text{再現率}\,(PN) &= \frac{\text{抽出正解用語数}\,(PN)}{NTCRIR\text{-1 の } TMREC \text{ テストコレクション中の全正解用語数}} \\
F \text{ 値} &= \frac{2 \times \text{再現率}\,(PN) \times \text{適合率}\,(PN)}{\text{再現率}\,(PN) + \text{適合率}\,(PN)}
\end{aligned}$$

候補語 3,000 語、6,000 語、9,000 語、12,000 語、15,000 語のそれぞれについて、抽出正解用語数、再現率、適合率、精度と再現率の調和平均である F 値の結果を示す。表 2.5 は抽出正解用語数であり、表 2.6 [15] は再現率、適合率、F 値の結果となる。

表 2.5: 抽出された完全一致用語数

PN	連接種類 LR	連接頻度 LR	FLR	語頻度	MC-value
3,000	1746	1784	1970	2034	2111
6,000	3270	3286	3456	3740	3671
9,000	4713	4744	4866	4834	4930
12,000	5974	6009	6090	5914	6046
15,000	7036	7042	7081	6955	7068

結果、スコア上位の候補、および 12,000 語以上を抽出する場合においては、湯本らが提案する FLR 法の性能が優れており、一方、1,500〜10,000 語程度の専門語を抽出したい

[13]NTCIR-1 で準備された用語である。

[14]NTCRIR-1 準備された形態素解析済みのコーパスから、MC-value 法、連接種類 LR 法、連接頻度 LR 法、語頻度法で定義された方法でスコア付けし、スコアの降順にソートを行い、こうして作られた用語候補を PN 個取り出した場合について、正解用語と比較する。

[15]表の各セルの内容は上段が再現率、中段が適合率、下段が F 値を表わす。

表 2.6: 抽出された完全一致用語数における再現率、適合率、F 値

PN	連接種類 LR	連接頻度 LR	FLR	語頻度	MC-value
3,000	0.197	0.202	0.223	0.230	0.239
	0.582	0.595	0.657	0.678	0.704
	0.295	0.301	0.333	0.343	0.356
6,000	0.370	0.372	0.391	0.423	0.415
	0.545	0.548	0.576	0.623	0.612
	0.441	0.443	0.466	0.504	0.496
9,000	0.533	0.536	0.550	0.547	0.557
	0.524	0.527	0.540	0.537	0.548
	0.529	0.532	0.545	0.542	0.553
12,000	0.676	0.680	0.689	0.669	0.684
	0.498	0.501	0.508	0.493	0.504
	0.573	0.577	0.584	0.567	0.580
15,000	0.796	0.796	0.800	0.786	0.799
	0.469	0.469	0.472	0.464	0.471
	0.590	0.591	0.594	0.583	0.593

のであるなら、MC-value 法の方が優れた結果を出したが、正解用語を含む長めの語でよいのであれば、FLR 法の語の出力は、正解用語の大部分をカバーすることができることを示した。

2.2.6 固有表現抽出の手法

文書の情報抽出を目的とした特徴語抽出に、固有表現抽出というタスクがある。固有表現抽出システムには大きく分けて、パターン照合規則を用いるものと、コーパスを用いた抽出規則の学習に基く手法がある。

パターン照合に基く固有表現抽出とは、人手で明示的なパターンを作成し、適用することによって固有表現を抽出するものである [17]。

例えば「○○君」、「○○円」、「株式会社○○」など、固有表現に含まれる接尾辞・接頭辞などの表記パターンを利用して文書中の固有表現を抽出する。パターン照合に基く手法の利点は、どのような規則が適用されているかを容易に観察することができ、限定された分野となるが、抽出精度が高くなることが期待できる点である。しかし、欠点としては、規則の変更や追加に対しては、常に人手で表記のパターンを作成することになり、多大なコストを要することが挙げられる。

現在では、人手で抽出規則を作成する手法に対し、固有表現のタグ付けされたコーパスから、機械学習を用いて抽出規則を自動的に学習する手法の研究が盛んにされている。機

械学習を用いる手法では、タグ付けされたコーパスが用意できれば、新たな固有表現に対しても抽出規則の生成に人手を要することがなく、再現性の高い規則を生成することが期待できる。

日本語固有表現抽出において用いられる機械学習の手法としては、決定木、最大エントロピー法、SVM などがある。固有表現抽出は入力文を適当な解析単位（トークン）に分割し、その単位にもとづき固有表現部分をまとめあげるという手法が一般的であり、トークンをまとめあげるタスク（チャンキング）に、山田ら [18] や Asahara ら [8] が SVM に基く手法を用いている。トークンの単位には単語や文字が考えられるが、Asahara らは、文字を用いた手法が単語を用いた手法よりも高い抽出精度が得られることを示した。しかし、それらの手法では、該当文字の前後 2 文字程度の品詞情報のみを用いてまとめあげを行うため、固有表現の構成単語数が多くなるにつれて、正確に抽出するのが困難になるという問題がある。

中野 [10] は、固有表現の抽出の手法として、文字単位だけでなく文節区切りまでも行い、文節内の情報を固有表現抽出のための素性として利用した。CRL 固有表現データ [10] を用いた評価実験の結果、F 値約 0.89 という結果を示し、提案手法の有効性を確認している。ここで、F 値は以下のとおりである。

$$
\begin{aligned}
\text{精度} &= P = \frac{\text{正しく抽出できた固有表現数}}{\text{システムが抽出した固有表現数}} \\
\text{再現率} &= R = \frac{\text{正しく抽出できた固有表現数}}{\text{データ中の全固有表現数}} \\
F\text{ 値} &= \frac{2 \times R \times P}{R + P}
\end{aligned}
$$

この提案されている特徴語抽出の詳細な手順を以下に説明する。機械学習に基づく固有表現抽出は、トークン系列を一つのまとまりとして同定し、そのまとまりに対して固有表現の種類を付与する処理である。すなわち、各トークンのまとめ上げと分類を組み合わせたタスクとみなすことができる。固有表現の開始、中間、終了など固有表現のまとめ上げ状態を表すタグを付与し、まとめ上げタスクと分類タスクを単一のタスクとして定式化することが可能になる。

この各トークンのまとめ上げ状態を表すタグとしては、先行研究として様々な手法が提案されている。中野は、Inside-Outside 法 [19] のバリエーションの一つであり、SVM を用いた固有表現抽出 [18] において最も精度が良いと報告されている、IOB2[20] と呼ばれる手法を用いている。IOB2 では固有表現の先頭トークンに B タグを付与し、それ以降のトークンに I タグを付与する。O タグは固有表現以外のトークンに付与される。このような下での固有表現抽出の学習は、対象トークンを中心とする前後 2 トークン前後の文脈を考え、そのトークンとそれに付属する品詞情報、文字種などの情報をベクトルにしたものを素性ベクトル x とし、対象トークンの固有表現タグ y との組 (x, y) を複数抽出し、機械学習のアルゴリズムにて、x から y の推定を行う分類器 $f(x)$ の学習をすることにな

る。そして未知の文書から素性ベクトルを生成し、学習された分類器より各トークンに対して固有表現タグを推定することになる。最終的に、固有表現部分は推定された固有表現タグから決定される。

先行研究の多くは、対象トークンの前後 2 文字などの固定長の文脈情報を用いており、この手法ではチャンキングの推定に必要な情報がチャンカーに与えられない場合がある。この問題に対処するためには、対象となっているトークンの窓外の情報を文脈に応じて適切に利用する必要があり、中野らの手法では、文節区切りを用いることによって、各文節の長さに応じた素性展開を行った。用いている文節の素性は以下の通りである。

1. 文節内素性
2. 隣接文節素性
3. 主辞素性

最初の文節内素性は、文節内で固有名詞が存在すれば、最も近い固有名詞の品詞細分類を用い、固有名詞が存在しなければ、文節の先頭の単語を素性として用いる。構成要素数が多い固有表現の内部には、地名や組織名が含まれることが多く、文節内素性を用いることによって、同一の素性が付与され、連続する名詞に対しての区切りの区別ができるようになる。

次の隣接文節素性は、解析方向に隣接する文節の末尾が名詞である場合に、その単語を素性として用いる。これは、一般的に文節は自立語と付属語から成り、文節が名詞で区切られている場合には何らかの重要な情報が含まれているとの考えからである。

最後の主辞素性は、各文節の主辞、つまり文節末から見て最初の自立語を素性とする。この素性により、地域や人名などが混合した複合名詞を正しく抽出することが期待できる。

中野の手法は、上記の素性を用いて素性展開後、分類器の学習を行い固有表現の抽出を行った。また、文節素性が抽出精度に与える影響を調べるために、以下に示す 4 種類の素性設定のモデルについて比較を行っている。[16]

1. base model: 文字, 単語, 品詞, 文字種, 予測対象のタグの $i+1, i+2$ 番目で予測された固有表現タグ
2. model A: base model の素性 + 文節内素性
3. model B: base model の素性 + 文節内素性 + 隣接文節素性
4. model C: base model の素性 + 文節内素性 + 隣接文節素性 + 主辞素性

[16] ここで文字種は、「ひらがな」、「カタカナ」、「アラビア数字」、「アルファベット小文字」、「アルファベット大文字」、「その他」の 6 種類を用いている。

評価実験にはCRL固有表現データ[17]を用いている。評価にはCRL固有表現データを記事単位に5等分し、訓練データ4、評価データ1の比率で交差検定を行い、それらの平均の精度、再現率、精度と再現率の調和平均であるF値で各モデルの比較を行っている。また、固有表現の構成要素数によって抽出精度がどのように変化するかの比較も行っている。

表2.7は、固有表現の種類ごとの精度の比較を示し、表2.8は、固有表現を構成する形態素数による抽出精度の変化を、表2.9は、抽出数および正解数の変化を示す。[18]

表 2.7: 固有表現の種類ごとの精度の比較

抽出タグ	頻度	base model	model A	model B	model C
ORGANIZATION	3676	80.46	84.12	84.31	84.30
PERSON	3840	87.88	88.88	89.08	89.16
LOCATION	5463	88.34	89.85	90.13	89.91
ARTIFACT	747	50.66	51.25	52.90	52.20
DATE	3567	94.55	94.71	94.73	94.68
TIME	502	89.29	89.49	91.45	91.24
MONEY	390	93.59	94.06	93.50	94.43
PERCENT	492	96.87	96.79	97.08	97.16
TOTAL	18677	87.07	88.50	88.78	88.72

表 2.8: 固有表現の構成要素数による精度

	構成する形態素数	$n \geqq 1$	$n = 1$	$n = 2$	$n = 3$	$n \geqq 4$
	頻度	18677	10349	5379	1689	1260
base model	F 値	87.09	88.81	89.51	76.51	74.31
	再現率	85.44	89.40	87.21	69.92	66.11
	精度	88.81	88.23	91.94	84.48	84.83
model C	F 値	88.72	89.86	91.17	80.58	79.00
	再現率	86.65	89.07	88.18	75.67	73.41
	精度	91.02	90.66	94.37	86.18	85.49

この結果、中野らは、F値約0.89という結果を得ている。また、表2.8より、F値の比較では、n が多くなるにしたがってbase modelとの精度の差が大きくなり、提案手法が構成要素数の多い固有表現に対しての有効性を示している。

[17]毎日新聞95年度版1,174記事、約11,000文に対してIREXで定義された固有表現が付与されている。
[18]表中の n は固有表現を構成する形態素数を示している。

表 2.9: 固有表現の構成要素数による抽出数と正解数

	構成する形態素数	$n \geqq 1$	$n = 1$	$n = 2$	$n = 3$	$n \geqq 4$
	頻度	18677	10349	5379	1689	1260
base model	抽出数	17968	10486	5102	1398	982
	正解数	15957	9252	4691	1181	833
model C	抽出数	17759	10168	5026	1483	1082
	正解数	16164	9218	4743	1278	925

この結果により、従来方法と比較して過抽出を抑えつつ、構成要素数の多い固有表現を抽出できるようになった事により、提案手法の有効性を示した。

2.3 各手法の考察

ここまでで、特徴語抽出の技術ついて先行研究がなされている様々な手法の紹介を行った。特徴語を抽出する目的として、

1. 情報検索 : 情報検索のため、文書を特徴付ける索引付けを行うことを目的とした手法
2. 情報抽出 : あらかじめ指定された情報を文書中から抽出することを目的とした手法
3. 専門用語抽出 : 専門分野のコーパスから専門用語を自動的に抽出することを目的とした手法

といった 3 つの目的に各特徴語抽出の手法の分類を行った。

これまでに紹介してきた先行研究における各特徴語抽出の手法をこのように分類してみると、表 2.10 のようになる。

表 2.10: 手法の目的

手法	目的
TF-IDF による手法	情報検索
SVM を利用する手法	情報検索
KeyGraph による手法	情報検索
χ^2 値を用いる手法	情報検索
出現頻度と連接頻度に基づく手法	専門用語抽出
固有表現抽出の手法	情報抽出

ここで実用的な特徴語抽出システムの構築について考えてみる。一般的に、特徴語抽出での教師データとなりうる大規模なコーパスまたは文書の統計データを用意するのは、多大なコストが掛ると考えられる。

いくらか、自然言語処理のためのリソースとして、京都大学テキストコーパス[19] や NAIST テキストコーパス[20] などが挙げられるが、特に実用的な業務での利用を想定すると、その業務での利用上での問題に沿ったコーパスや文書の統計データが必要となってくる。また、大量の文書データが入手できたとしても、その頻度などの統計情報を集計計算するだけであれば、機械的に処理を行うことは容易であるが、その文書データに対して正解となるラベルを付けていくことを、機械的に行うのは困難である。よって、人手でラベルを付けていくことになり、多大な時間とコストが掛ってしまう。

[19] http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス
[20] http://cl.naist.jp/nldata/corpus/

このような理由から、これまでに説明した特徴語抽出の手法を、大量のコーパスや統計データの必要性を背景とした手法であるかどうか、また、構築のコストやスケーラビリティの観点から特徴語抽出の手法の考察を行う。

まず、特徴語抽出にあたり、大規模なコーパスや文書の統計データが必要かどうかという視点から分類を行う。大規模なコーパスまたは統計データが必要な手法として、「TF-IDF による手法」、「SVM を利用する手法」、「出現頻度と連接頻度に基づく手法」、「固有表現抽出の手法」として分類され、大規模なコーパスまたは統計データを必要としない手法としては、「KeyGraph による手法」「 χ^2 値を用いる手法」として分類することができる。

TF-IDF による手法は、特徴語抽出の対象となる文書集合が与えられた時点で計算できることになり、特徴語抽出の精度は与えられた文書集合の量に依存することになるが、大規模な文書の統計データを入手することができれば、一定の精度を期待することができる。また、ほぼ集計処理のみの計算コストで済み、また実装も容易なことから、特徴語抽出などの手法の比較のためのベースラインとして利用される。

また、出現頻度と連接頻度に基づく手法に関しても、大規模な文書の統計データを入手することができれば、特徴語抽出の計算を行うことができるが、単純な文書の統計処理に加えて、単名詞のバイグラムの集計計算を行い、単名詞の左右に連接する語の種類数あるいは頻度の集計計算が必要であり、さらには、単名詞を複合名詞に拡張したスコアの計算が必要となってくる。よって、TF-IDF による手法と比べて計算コストが高くなってしまう。

また、大規模なコーパスが必要である SVM を利用する手法や、固有表現抽出の手法については、教師あり学習のアルゴリズムを利用することになり、事前にラベル付きの正解データを用意しなければならない。

例えば、SVM を利用する手法では、特徴語候補集合から、それらの語を特徴語となるかどうかを判断しなければならない。この判断は対象文書集合での特徴語かどうかの問題に依存しており、機械的に判断することは難しく、人手によって特徴語かどうかの判断を行うことになる。固有表現抽出の手法に関しても、解析単位（トークン）を固有表現としてまとめあげるタスク（チャンキング）で SVM を利用しており、各トークンのまとめあげ状態を表すタグを人手によって教師データとなる文書に付与していくことになる。

このように、人手によって特徴語かどうかの判断や、トークンのまとめあげ状態などの正解を、教師データに対してラベル付けを行わなければならず、多大なコストと時間が掛ってしまうことになる。また、人手で正解のラベルを与えることになり、その作業をする人の主観に依存してしまう。

処理時間に関して述べれば、一般的に SVM は特徴量の次元数が大きくても対応できるが、教師データ数が増加してくれば学習の処理時間とデータ数の関係は線形ではなくなってくる。特に SVM を利用する手法では、SVM のカーネル関数にガウシアン型カーネル [3] を用いている。分類での未知データに対しての汎化能力は期待できるが、学習に膨大な時間が掛ってしまうことが想像される。実用的な特徴語抽出システムを考えた場合、このような教師あり学習のアルゴリズムを利用するにあたり、適時に新しい教師データの獲

得が期待できる場合には、その教師データを使って学習モデルを更新し、分類精度の向上に努めるのが望ましい。しかし、基本的にはSVMはバッチ学習型のアルゴリズムであることから、膨大な時間が掛る学習を随時行うことは難しいといえる。

カーネル関数を線形カーネルに限定し、SGDアルゴリズムで学習するような、オンラインSVMと呼ばれるアルゴリズムが考案されている。SVMに比べて精度面で若干劣るものの、高速に大規模データを処理できるという利点から近年盛んに利用されているアルゴリズムである。このSVMを利用する手法の実用的なシステム構築を考えた場合、オンライン学習型のSVMのアルゴリズムの導入を考慮するのが望ましいといえる。

次にKeyGraphによる手法と χ^2 値を用いる手法に関して、大きな特徴のひとつとしては、大規模なコーパスや文書の統計データを必要としない手軽さにある。コーパスを利用することなく、単一の文書だけから特徴語を抽出するには、語の出現頻度を用いる方法[11]や「要するに」などの手がかり語をもとに特徴語を抽出する方法[21, 22]などがある。しかし、語の出現頻度を用いる方法は単純すぎて、一般的な語も抽出してしまい、また、手がかり語をもとに特徴語を抽出する方法は汎用性がない。KeyGraphによる手法と χ^2 値を用いる手法は、対象とする文書だけの情報から、語の共起をもとに統計的な指標を用いて特徴語を抽出する手法である。両手法ともTF-IDFをベースラインに比較を行っており、被験者に特徴語であるかどうかを判定してもらう評価方法であるが、TF-IDFに匹敵する性能が得られている。

処理時間に関してもKeyGraphによる手法は、対象文書の語数を W とした場合に、処理時間は W に対してわずかに速く増大したほどであり、χ^2 値を用いる手法では、ほぼ語数に対して線形なオーダで処理が終了している。

また、 χ^2 値を用いる手法は、KeyGraphによる手法が基礎になっており関連が深い。KeyGraphによる手法では、語と土台中の語の共起度で計算したkey値を用いて語の重み付けを行う。key値が高くなるためには、特定の土台と偏って共起することが必要であるので、意味するところは χ^2 値を用いる手法のアイディアに近い。よって、 χ^2 値を用いる手法は、KeyGraphによる手法を統計的に洗練した手法であるとも考えることができる。

また、本研究で行う実験は、特に法律文の検索システムへの利用を想定しており、情報検索のため、文書を特徴付ける索引付けを行うことを目的とした、特徴語抽出の手法について焦点を置くこととする。

ここまでの考察から、実用的な特徴語抽出システムの構築を想定した場合に、事前に用意するコーパスや文書の統計データの収集のコストが低く済むことと、特徴語抽出に要する処理時間がシステム的に許容できる時間内であることが重要と考える。よって、次章にて上記の条件を兼ね備えている χ^2 値を用いる手法を用い、実際の法律文に対して特徴語抽出の実験を行うこととする。

表 2.11: 手法の分類

手法	データ準備コスト	計算コスト
TF-IDF による手法	中：文書の統計データが必要	低：集計処理のみの計算コスト
SVM を利用する手法	高：正解ラベル付きコーパスが必要	高：学習の計算コスト
KeyGraph による手法	低：事前データは必要なし	中：語数に対してほぼ線形な計算コスト
χ^2 値を用いる手法	低：事前データは必要なし	低：語数に対して線形な計算コスト
出現頻度と連接頻度に基づく手法	中：文書の統計データが必要	中：複雑な集計処理の計算コスト
固有表現抽出の手法	高：正解ラベル付きコーパスが必要	高：学習の計算コスト

第3章　実験

3.1　実験内容について

冒頭で述べたように、依然法律文の特徴語の辞書化やタグ付けなどの資源の整備が進んでいると言えず、法律という領域において、特徴語抽出の手法の評価を行うことは重要であると考えられる。そこで本研究において、特徴語抽出の先行研究の手法を用いて、実際の法律文に適用して実験を行うこととする。

これまでの考察を踏まえ、実用的な特徴語抽出システムの構築を想定した場合に、事前に用意するコーパスや文書の統計データの収集のコストが低く済むことと、特徴語抽出に要する処理時間がシステム的に許容できる時間内であることが重要であると考えられる。

特徴語抽出において様々な手法が提案されているが、その中でも、上記の条件を兼ね備えている χ^2 値を用いる手法が適していると考えた。そこで本研究では、 χ^2 値を用いる手法を実際の法律文に適用し、実験を行うこととする。また、特徴語抽出の一般的な目安として、語の出現頻度との比較を行い、問題点を洗い出し考察を行う。

本研究で対象とした法律文は、複数の法律分野での特徴語抽出の差を比較するために3つの法律分野から法律文を選択し、特徴語抽出の実験を行うこととする。

それぞれ、以下のように、社会保険分野、労働分野、建築住宅分野から法律文の選択を行った。

1. 社会保険分野：国民年金法（平成 23 年 5 月 1 日現在の法令データ）

2. 労働分野：労働基準法（平成 23 年 5 月 1 日現在の法令データ）

3. 建築住宅分野：建築基準法（平成 23 年 5 月 1 日現在の法令データ）

これらのデータは、「法なび法令検索」[1] で提供されているデータを収集し、データベースへと保存する。本研究で検証する特徴語抽出は、図 3.1 のように行う。

本研究では法律文の集合を対象としており、その中に含まれるテキストを形態素解析する。

分割された形態素は、複合語として連結できるものもあるため、これらを連結する。こうして特徴語候補のもととなる語の集合から、いくつかの制約によるフィルタリングを行い、特徴語候補を抽出する。

[1] http://hourei.hounavi.jp/

抽出された特徴語候補より、頻出語の選択を行い、選択された頻出語と特徴語候補との共起頻度の集計を行う。得られた語の集計データより、χ^2 値の計算を行い、特徴語を抽出する。

3.2 法律文における特徴語抽出実験

対象となる法律文は日本語であり、日本語文は、英文と違い単語間の明確な区切りが存在しない。そのため、文書中から語を取り出すためには、形態素解析と呼ばれる処理を施さなければいけない。形態素解析とは、テキストを形態素と呼ばれる単位に分割することを指す。この形態素は、厳密にいえば単語とは違った分割の単位ではあるが、おおよそ単語と同じようなものになる。そのため、形態素は品詞の情報を持つ。例えば以下のような文を形態素解析した場合、表 3.1 のような形態素に分割される。

被保険者は、保険料を納付する。

表 3.1: 形態素解析の例

形態素	品詞
被	接頭詞, 名詞接続
保険	名詞, 一般
者	名詞, 接尾, 一般
は	助詞, 係助詞
、	記号, 読点
保険	名詞, 一般
料	名詞, 接尾, 一般
を	助詞, 格助詞, 一般
納付	名詞, サ変接続
する	動詞, 自立
。	記号, 句点

このように、形態素解析による文書解析により、日本語の法律文の中から語を取り出すことができる。形態素解析のプログラムとして、Mecab [2] や ChaSen [3] といったものがある。表 3.1 の形態素解析の結果は、Mecab による解析の結果である。Mecab と ChaSen では、形態素の品詞体系が異なるため、解析の結果が大きく変わる。本研究では、品詞の情報を多く用いるため、より詳細な品詞の分類が可能な Mecab を用いて形態素解析を行う。また、Mecab は辞書を利用することになり、辞書として IPA 辞書[4] を用いることにする。これは、IPA コーパスに基づき CRF [5] でパラメータ推定した辞書になる。

形態素解析により得られた形態素は、語を構成する最小単位と考えることができる。したがって、この形態素の連結により新たな語となることがある。そこで、特徴語候補のもとになる語の集合を作るために、形態素の連結処理を行う。

[2]http://mecab.sourceforge.net/
[3]http://chasen-legacy.sourceforge.jp/
[4]http://mecab.sourceforge.net/src
[5]http://www.cis.upenn.edu/ pereira/papers/crf.pdf

連結して新たな語となるかどうかは、その形態素の品詞により判別する。表 3.1 の形態素の中から連接することができるものを考えると、「保険」と「料」は連結により「保険料」になると考えられる。また「納付」と「する」を連結し「納付する」とすることもできると考えられる。この場合「保険料」は連結により語とする意義はあるが、「納付する」に関してはその意義は薄いと考える。

「保険」、「料」はともに名詞である。このように名詞が連続する場合は、連結することにより新たな語となることがある。しかし、「納付」と「する」は名詞と動詞の連接である。この場合は「納付」の動詞化となるため、特徴語として抽出する意義は薄い。また、形態素解析は利用する辞書などに解析結果が依存してしまうため、常に正確な分割を行うとは限らない。例えば、「国民年金」という単語は「国民」と「年金」に分割されてしまい、この二つの品詞も名詞となる。そこで、本研究では連結処理を行うのは名詞が連接している場合とする。これにより、複合語などの二つ以上の形態素からなる語を特徴語の候補として選ぶことができるようになる。

このようにして得られた語の集合にも、特徴語となりえない語は含まれている。形態素の品詞は、名詞にもいくつかの小分類が存在する。この小分類の中で、特徴語の先頭に来ることのないような項目が存在する。それらは、「非自立」、「接尾」、「代名詞」である。非自立の品詞情報を持つ形態素は、「こと」、「もの」のように単独では意味をなさない語であり、接尾の品詞情報を持つ形態素としては「的」や「化」といったものである。これらの形態素が語頭に来る語は、そもそも語としての条件を満たしていないと考えられる。これらの形態素も、語頭に来ることはないと考えられるが、代名詞に関しては、語頭となることはあり得る。しかし、代名詞に連接する語は、「この条」、「あの規定」などのように、何かを指示するときに使用する。したがって、これらの語は特徴語として考慮する必要はない。このようなことから、先頭の形態素として「接尾」、「非自立」、「代名詞」が来ている語は、特徴語候補には含めないこととする。

以上のように特徴語候補となる集合が得られた。次に頻出語の選択を行う。これは、特徴語候補を抽出した時と同じく法律文に対して形態素解析を行い、文書中の語の延べ総数の 30%に達するまで頻出語の上位語を取り出す。この得られた頻出語の集合から文書自身の全体的な傾向が伺えることになる。国民年金法、労働基準法、建築基準法から得られた頻出語の上位語は以下のようになる。

それぞれ、表 3.2 は国民年金法から抽出した頻出語上位 50 語であり、表 3.3 は労働基準法から抽出した頻出語上位 50 語、表 3.4 は建築基準法から抽出した頻出語上位 50 語である。

この文書自体の傾向から大きく逸脱する特徴を持つ語を特徴語として取り出すために、頻出語と特徴語候補の共起頻度を量ることになるが、この 2 語間の共起の定義を文単位として共起頻度の集計を行う。これは、共起度を文書中で連続する長さ W の範囲中で 2 語両方が出現する頻度とすると、 W が 1 文より短い場合には、疑問形や倒置によって語の順序が変わると、関連があると思われる語同士が離れることになり、共起していると判断できなくなってしまうことがある。また反対に、 W を 1 文より長くすると、指示語の形

などで複数文に跨り出現しても共起と捉える事ができるが、実際には意味のつながりの薄い語の対の方が多く、共起していると判断するべきでないとの考えからである。

次に、頻出語と共起する語の χ^2 値を計算する。頻出語単独での生起確率を理論確率 $p_g(g \in G)$ とし、語 w と頻出語群 G の共起の総数を n_w 、語 w と語 $g \in G$ の共起頻度を $freq(w, g)$ とすると、統計量 χ^2 値は以下の式 3.1 で与えられる。

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w, g) - n_w p_g)^2}{n_w p_g} \tag{3.1}$$

しかし、文書中の一文の長さは一様ではなく、長い文に出現する語は他の語と共起する確率が高まり、逆に短い文に出現する語は共起する確率が低くなる。語の共起の定義を文単位としているので、共起の確率は文の長さに依存することになる。逆に短い文において共起している時は、より 2 語の関連が強いと考えることができる。このような考えより、統計量 χ^2 値に対して次のような改良を行う。

1. p_g を (g が出現する文の語数の合計) / (文書全体の語数の合計) とする。

2. n_w を語 w が出現する文の語数の合計とする。

このように、文書中の任意の語が g と共起している確率 p_g に語 w と共起する語の合計数 n_w を乗じて共起の期待頻度とすることにより、文の長さを考慮した結果が期待できる。また、頻出語の中の特定の語 g とだけ共起する語は、語 g に付随する語である場合が多く、特徴的な語では無いにも拘わらず χ^2 値が高くなってしまう。このような分布の偏りを防ぐ目的で、式 3.2 にて χ^2 値の最大の項を除く工夫を施す。

$$\hat{\chi^2}(w) = \chi^2(w) - \max_{g \in G} \frac{(freq(w, g) - n_w p_g)^2}{n_w p_g} \tag{3.2}$$

このように、すべての特徴語候補中の語 w について、頻出語との共起頻度と w が出現する語の総数を集計し χ^2 値の計算を行う。得られた語の χ^2 値より、降順に一定数の語を特徴語として提示する。

3.3 実験の評価

本実験の結果を以下に示す。表 3.5 は国民年金法に対して本実験の手法を適用した結果として、特徴語として抽出された上位 50 の語である。また、表 3.6 は労働基準法に対しての結果であり、表 3.7 は建築基準法に対して適用した結果となる。[6]

抽出された法律文の特徴語として、国民年金法からは「養子」、「母子福祉年金」などの語が、労働基準法からは「従前」、「機関」などの語が、建築基準法からは「資格」、「公益」などの語が上位になっていることがわかる。

文書の特徴語の妥当性の判断は、明確な基準を設けるのは難しく、筆者の感覚的な判断により、あくまで主観的な評価になってしまうが、国民年金法は特に「福祉年金」関連の語を特徴的と捉え、労働基準法は「省庁」関連の語を、建築基準法に関しては「建築構造」関連の語を特徴的と捉えていることが窺える。これらは、法律文の一部分の概念が特定された特徴的な語が抽出されているように見受けられる。

また、それぞれの頻出語の上位語と比較を行ってみる。頻出語の上位語においては、国民年金法では「期間」、「法律」、労働基準法、建築基準法は供に「法律」、「基準」など法律文においては当たり前の語が上位になっているのに対し、χ^2 値の上位語においては、国民年金法では「福祉年金」、「罰則」、労働基準法は「職員」、「事務」、建築基準法では「特例容積」、「延べ面積」など比較的出現頻度が低い語であっても、それぞれの法律において特徴的であると考えられる語を取り出している。

このように、本実験で用いた χ^2 値を用いる共起情報に基づく手法は、考案者である松尾らも述べているように、頻出語を基準とするが、出現頻度による頻出語で、すでに十分よい特徴語になっている場合には、文書の一部分の概念が特定された特徴的な語になっているケースが多く、逆に頻出語が「月」や「例」など一般的な語が多く、特徴語としての情報量が少ない場合には、適切な特徴語となっているケースが多くなる傾向が窺えた。

特に法律文においては専門的な語の出現が多く、頻出語で十分よい特徴語となっており、法律文の一部分の概念が特定された語が抽出される結果になったと考えられる。

筆者の主観的な判断となるが、これらの実験結果の特徴語上位 50 語に対して、特徴語としてふさわしいかどうかの判定を行った。実験結果を表 3.8 に示す。

国民年金法では、正解率 34%となり、労働基準法では、正解率 28%、建築基準法では、正解率 40%という結果となった。

また、国民年金法で述べれば、語の出現頻度が高く、特徴語であると考えられる「国民年金」、「厚生年金保険」などが比較的下位の順序となっている。これは、法律文の内容として「国民年金」などは頻出語である「保険」、「法律」などの語と共起することが多く、こうした頻出語との共起関係が一様な分布となりやすく、頻出語集合との間に何らかの意味的なつながりとなる共起関係の有意性が現れなかったことが原因として考えられる。

このように、特徴語抽出における精度という観点からは、それぞれの法律文に対して全

[6]表 3.5, 表 3.6, 表 3.7 の「判定」項目は、筆者の主観により特徴語であると判定した語に対して「○」を付けてある。

般的に正解率が低い結果となっている。実用的な特徴語抽出システムの構築を想定した場合では、特徴語としての精度が期待できない。よって、この手法をそのまま用いるのは難しいと考えられ、何かしらの精度向上の改良が必要であると思われる。

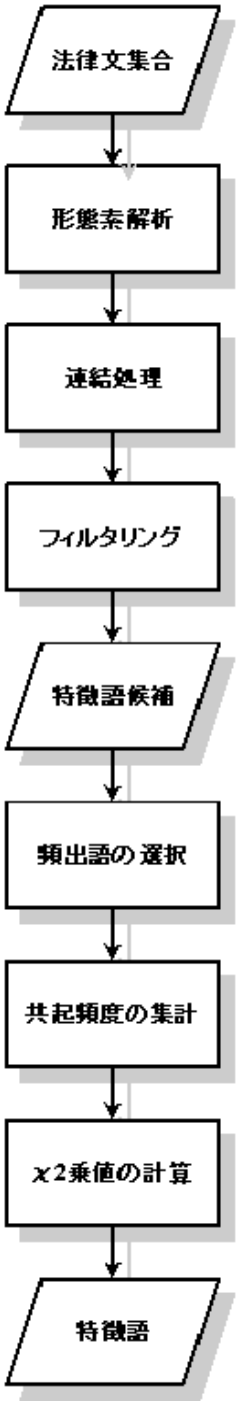

図 3.1: 特徴語抽出の流れ

表 3.2: 国民年金法の頻出語の上位語

順位	ラベル	頻度	順位	ラベル	頻度
1	保険	1290	26	障害	155
2	期間	846	27	障害基礎年金	129
3	法律	701	28	従前	127
4	国民年金	681	29	状態	127
5	額	572	30	程度	115
6	昭和	494	31	老齢年金	110
7	年金	425	32	前条	108
8	附則	363	33	老齢基礎年金	98
9	法	311	34	資格	97
10	月	281	35	省令	90
11	条	278	36	端数	90
12	前項	271	37	傷病	90
13	政令	250	38	間	89
14	厚生	247	39	妻	88
15	項	244	40	事項	86
16	基金	185	41	事由	86
17	次	177	42	つて	84
18	月数	174	43	初診	83
19	厚生年金保険	170	44	権利	78
20	平成	167	45	理事	78
21	大臣	165	46	共済組合	78
22	母子福祉年金	164	47	廃疾	78
23	なつ	159	48	夫	77
24	子	157	49	表	76
25	例	157	50	至	74

表 3.3: 労働基準法の頻出語の上位語

順位	ラベル	頻度	順位	ラベル	頻度
1	法律	255	26	定め	28
2	基準	133	27	間	27
3	期間	90	28	新法	25
4	事業	86	29	罰則	25
5	厚生	74	30	女性	24
6	賃金	65	31	日数	23
7	省令	63	32	昭和	22
8	条	63	33	委員	21
9	過半数	58	34	有給休暇	21
10	例	58	35	書	19
11	事項	56	36	書面	18
12	前項	56	37	災害	18
13	組合	54	38	局長	17
14	附則	51	39	事由	17
15	項	46	40	条件	15
16	従前	45	41	規則	15
17	業務	43	42	行政	14
18	政令	41	43	法	14
19	行政官庁	39	44	福祉	14
20	部分	37	45	児童	14
21	前条	35	46	寄宿舎	13
22	範囲	34	47	割増賃金	13
23	つて	34	48	都道府県	13
24	平成	34	49	大臣	13
25	次	30	50	効力	12

表 3.4: 建築基準法の頻出語の上位語

順位	ラベル	頻度	順位	ラベル	頻度
1	法律	332	26	事項	82
2	基準	316	27	防火	79
3	区域	252	28	事務	77
4	構造	243	29	条例	74
5	政令	226	30	範囲	69
6	部分	218	31	支障	62
7	行政	208	32	期間	60
8	前項	208	33	容積	59
9	地区	194	34	衛生	59
10	国土交通大臣	190	35	規模	57
11	項	188	36	従前	57
12	敷地	172	37	附則	54
13	機関	168	38	道路	54
14	業務	162	39	例	52
15	国土交通省令	149	40	対象区域	50
16	主事	134	41	別表	49
17	土地	133	42	用途	49
18	都市	132	43	技術	45
19	条	132	44	限度	45
20	つて	120	45	防火地域	42
21	次	119	46	壁面	42
22	資格	119	47	都道府県	41
23	市町村	108	48	型式部材	41
24	都道府県知事	96	49	おそれ	40
25	前条	83	50	職員	40

表 3.5: 国民年金法の χ^2 値の上位語

順位	ラベル	χ^2 値	頻度	判定	順位	ラベル	χ^2 値	頻度	判定
1	理事	21.13	78		26	機関	5.68	16	
2	養子	21.06	59	○	27	定め	5.60	16	
3	母子福祉年金	14.64	164	○	28	直系姻族	5.12	12	○
4	業務	13.53	51		29	直系血族	5.12	12	○
5	基金	13.15	185		30	対象期間	5.11	45	
6	限度	12.80	31		31	祖母	5.03	19	○
7	規約	10.65	27		32	姉	5.03	19	○
8	省令	9.90	90		33	役員	4.99	23	
9	罰則	9.71	22		34	義務	4.87	24	
10	権限	8.34	56		35	自己	4.80	6	
11	法別表	8.30	30		36	状況	4.79	17	
12	事項	8.27	86		37	情報	4.78	15	
13	孫	8.14	41	○	38	使用人	4.70	6	○
14	行政	8.06	20		39	代理人	4.70	6	○
15	弟妹	7.72	41	○	40	委員	4.68	12	
16	本文	7.68	27		41	物価指数	4.65	70	○
17	事務	7.66	71		42	書類	4.54	9	
18	福祉年金	6.98	57	○	43	日本私立学校	4.51	16	○
19	総会	6.73	20		44	共済事業	4.51	16	○
20	機構	6.33	37		45	目的	4.42	16	
21	代議員	6.27	28	○	46	理由	4.35	20	
22	資格	6.27	97		47	金融機関	4.31	9	○
23	事業	5.92	38		48	団体	4.24	16	
24	廃疾	5.90	78		49	事故	4.22	6	
25	老齢福祉年金	5.75	72	○	50	監事	4.19	16	

表 3.6: 労働基準法の χ^2 値の上位語

順位	ラベル	χ^2 値	頻度	判定	順位	ラベル	χ^2 値	頻度	判定
1	従前	8.26	45		26	組合	2.16	54	○
2	附則	6.32	51		27	地方公共団体	2.15	9	○
3	政令	6.21	41		28	賃金	2.07	65	○
4	罰則	4.68	25		29	辞令	1.85	2	○
5	機関	3.77	10		30	公社	1.81	4	
6	局長	3.15	17	○	31	ガス	1.80	3	
7	事務	3.14	10		32	場所	1.78	6	
8	職員	3.12	12		33	内閣	1.76	3	
9	法律	2.87	255		34	会長	1.74	3	
10	年数	2.87	8		35	昭和	1.72	22	
11	上欄	2.87	3		36	平成	1.69	34	
12	欄	2.87	5		37	府	1.66	4	
13	例	2.80	58		38	事業	1.65	86	
14	相当	2.75	6		39	郵政民営	1.63	2	○
15	農林水産省	2.72	2	○	40	別表	1.62	10	
16	外務省	2.72	2	○	41	厚生労働省	1.62	6	○
17	法務省	2.72	2	○	42	対象業務	1.61	10	
18	都道府県	2.57	13		43	地方自治	1.61	6	○
19	国	2.50	7		44	支払	1.60	3	
20	過半数	2.40	58		45	女性	1.58	24	
21	行政	2.35	14		46	児童	1.57	14	
22	表	2.28	11		47	銀行	1.54	1	
23	字句	2.27	4		48	郵便振替	1.54	1	○
24	申立て	2.18	12		49	郵便	1.54	1	○
25	効力	2.18	12		50	郵便為替	1.54	1	○

表 3.7: 建築基準法の χ^2 値の上位語

順位	ラベル	χ^2 値	頻度	判定	順位	ラベル	χ^2 値	頻度	判定
1	資格	6.18	119		26	型式部材	3.19	41	○
2	公益	5.70	28		27	帳簿	3.15	14	○
3	工業	5.20	16	○	28	期限	3.12	21	
4	業務	5.18	162		29	床	3.08	14	○
5	長	5.12	37		30	技術	3.06	45	
6	役員	4.88	19		31	開口	3.05	11	○
7	罰則	4.84	39		32	最高限度	3.04	31	
8	利便	4.61	16		33	附則	3.04	54	
9	機関	4.34	168		34	数値	3.00	37	
10	居室	4.21	24	○	35	床面積	2.99	19	○
11	理由	4.18	30		36	間	2.98	23	
12	特例容積	3.99	20	○	37	地域	2.98	37	
13	延べ面積	3.95	30	○	38	書類	2.91	23	
14	法人	3.92	24		39	名称	2.90	10	
15	事務所	3.76	27		40	住居地域	2.89	23	○
16	部分	3.62	218		41	階段	2.86	12	○
17	構造方法	3.59	38	○	42	地区	2.86	194	
18	基本方針	3.56	6		43	近隣商業地域	2.81	7	○
19	容積	3.55	59	○	44	新法	2.78	28	
20	工業地域	3.53	13	○	45	柱	2.77	18	○
21	壁	3.46	21		46	敷地面積	2.69	39	○
22	商業地域	3.46	8	○	47	高層住居	2.68	14	○
23	都道府県知事	3.35	96		48	行	2.67	20	
24	職員	3.29	40		49	線	2.66	9	
25	従前	3.22	57		50	主事	2.62	134	○

表 3.8: 特徴語抽出の結果

法律文	χ^2 値上位 50 語の正解数	精度
国民年金法	17	0.34
労働基準法	14	0.28
建築基準法	20	0.4

第4章　おわりに

近年、ますます多くの電子的な文書が蓄えられるにしたがって、その文書の内容を大まかに把握する目的で、文書の特徴語を抽出することは重要になってきている。その方法として、機械的に特徴語を抽出することができれば、事前の情報収集に大きく役立つことになり、機械的な特徴語抽出の手法に関して調査・考察することは重要であることを述べた。また、本研究では、情報検索を目的とした特徴語抽出をはじめとする先行研究を紹介した。

調査と考察を行い、実際のシステムに適用するにあたり、松尾らの χ^2 値を用いる手法 [5] が、事前に大量のコーパスや文書に関する統計データが必要ではなく、人手やコストが掛らず、また、スケーラビリティの観点からも有効であると考えた。

また、法律文の特徴語の辞書化やタグ付けなどの資源の整備が進んでいないことから、実際の法律文に松尾らの χ^2 値を用いる手法を適用し、実験・考察を行った。

特徴語抽出における精度という観点からは、実験を行った、それぞれの法律文に対して全般的に正解率が低い結果となっており、実用的な特徴語抽出システムの構築を想定した場合において、この手法をそのまま用いるのは難しい結果となった。

しかしながら、この手法は、事前にコーパスや文書に関する統計データを用意することなく、手元にある文書だけで処理できるという手軽さが大きな利点であり、この利点を活かした改良を行うことが重要である。例えば、単純に頻出語の集合との共起の統計量を用いるのではなく、対象となる文書の特性に沿った語の集合との共起の統計量を用いること、などが考えられる。

以上のように、本実験に用いた χ^2 値を用いる手法は、他の TF-IDF などの他の手法に置き換わるような手法ではないが、他の手法と組み合わせて用いることで、より適切な特徴語の抽出ができると考えられる。また、この利点を活かした改良を行うことが今後の課題といえる。

参考文献

[1] G. Salton and C.Buckley, Term-weighting approaches in automatic text rerieval, Infomation Processing and Management, Vol14, No.5, pp.513-523, 1988.

[2] 杉浦 広和, 議事録集合からの特徴語抽出とその応用に関する研究, 2009.

[3] 小野田 崇, サポートベクターマシン (知の科学), オーム社, 2007.

[4] 大澤 幸生, ネルス E. ベンソン, 谷内田 正彦, KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会誌, Vol.J82-D-I, No.2, pp.391-400, 1999.

[5] 松尾 豊, 石塚 満, 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌 17 巻 3 号 D, pp.217-223, 2002.

[6] 湯本 紘彰, 中川 裕志, 森 辰則, 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol.10, No.1, pp.27-46, 2003.

[7] Frantzi, K and Ananiadou, The C-value/NC-value method for ATR, Journal of NLP, pp.145-179, 1999.

[8] Masayuki Asahara and Yuji Matsumoto, Japanese Named Entity Extraction with Redundant Morphological Analysis, In Proc, HLTNAACL 2003, 2003.

[9] 中野桂吾, 日本語固有表現抽出における文節情報の利用, 2004.

[10] IREX 実行委員会（編）, IREX ワークショップ予稿集, 1999.

[11] Luhn H.P, A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and Development, Vol.1, No.4, pp.390-317, 1957.

[12] Sparck Jones, A Statistical Interpretation of Term Specificity and Its Application in Retrieval, Journal of Documentation, pp.11-21, 1972.

[13] K.W. Church and P. Hanks, Word association norms mutual infomation and lexicography, Computational Lingustics, Vol6, No.1, 1997.

[14] 東京大学教養学部統計学教室（編）, 統計学入門, 東京大学出版会, 1991.

[15] Nakagawa, H. and Mori, Nested Collocation and Compound Noun for Term Recogniton, In Proceedings of the First Workshop on Computational Terminology(COMPTERM 98), pp.64-70, 1998.

[16] Frantzi,K.and Ananiadou, Extraction Nested Collocations, In Proceedings of the 16th International Conference on Computational Linduistics (COLINFG 96), pp.41-46, 1996.

[17] 竹本義美, 福島俊一, 山田洋志, 辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出, 情報処理学会論文誌, Vol.42, No.6, pp.1580-1591, 2001.

[18] 山田寛泰, 工藤拓, 松本裕治, Support Vector Machine を用いた日本語固有表現抽出, 情報処理学会論文誌, Vol.43, No.1, pp.44-53, 2002.

[19] L.A.Ramshaw, M.P.Marcus, Text Chunking Using Transformation-based Learning, In Proc, the Third Workshop on Very Large Corpora (WVLC-95), pp.82-94, 1995.

[20] E.F.T.K.Sang, Noun Phrase Recognition by System Combination, In Proc, NAACL00, pp.50-55, 2000.

[21] Edmundson.H, New Methods in Automatic Abstracting, Journal of ACM, Vol.16, No.2, pp.264-285, 1969.

[22] 木本晴夫, 日本語新聞記事からのキーワード自動抽出と重要度評価, 電子情報通信学会誌, Vol.74-D-I, No.8, pp.556-266, 1991.

[23] 相澤 彰子, テキストコーパスにおける特徴語抽出のための分析ツール, 情報処理学会研究報告 情報学基礎研究会報告, Vol.2001, No.20(20010305), pp.113-120, 2001.

[24] 長尾真, 水谷幹男, 池田浩之, 日本語文献における専門用語の自動抽出, 情報処理学会論文誌, Vol.17, No.2, pp.110-117, 1976.

[25] 大平, 帆足, 松本, 橋本, 白井, AIC を用いた重要語抽出手法と重要語を用いたターム重みづけ手法の提案・評価, 知識発見のための自然言語処理シンポジウム, 1999.

[26] Makoto IWAYAMA, Takenobu TOKUNAGA, A Probablistic Model for Text Categorization: Based on a Single Random Variable with Multiple Values, Proceedings of 4th Conference on Applied Natural Language Processing, pp.162-167, 1994.

[27] H.P.Luhn, A staistical approach to mechanized encodiong and searching of literary information, IBM J.Research and Development, Vol1, No.4, pp.309-371, 1957.

[28] 西野文人, 日本語テキスト分類における特徴素抽出, 情処学研報, NL112, pp.95-102, 1996.

[29] 吉村 賢治, 日高 達, 吉田 将 (福岡大工), 日本語科学技術文における専門用語の自動抽出システム, 情報処理学会論文誌, Vol.27, No.1, pp.33-40, 1986.

付録

実験での χ^2 値一覧

表 4.1: 国民年金法の χ^2 値

順位	ラベル	χ^2 値	頻度	順位	ラベル	χ^2 値	頻度
1	理事	21.13	78	2	養子	21.06	59
3	母子福祉年金	14.64	164	4	業務	13.53	51
5	基金	13.15	185	6	限度	12.80	31
7	規約	10.65	27	8	省令	9.90	90
9	罰則	9.71	22	10	権限	8.34	56
11	法別表	8.30	30	12	事項	8.27	86
13	孫	8.14	41	14	行政	8.06	20
15	弟妹	7.72	41	16	本文	7.68	27
17	事務	7.66	71	18	福祉年金	6.98	57
19	総会	6.73	20	20	機構	6.33	37
21	代議員	6.27	28	22	資格	6.27	97
23	事業	5.92	38	24	廃疾	5.90	78
25	老齢福祉年金	5.75	72	26	機関	5.68	16
27	定め	5.60	16	28	直系姻族	5.12	12
29	直系血族	5.12	12	30	対象期間	5.11	45
31	祖母	5.03	19	32	姉	5.03	19
33	役員	4.99	23	34	義務	4.87	24
35	自己	4.80	6	36	状況	4.79	17
37	情報	4.78	15	38	使用人	4.70	6
39	代理人	4.70	6	40	委員	4.68	12
41	物価指数	4.65	70	42	書類	4.54	9
43	日本私立学校	4.51	16	44	共済事業	4.51	16
45	目的	4.42	16	46	理由	4.35	20
47	金融機関	4.31	9	48	団体	4.24	16
49	事故	4.22	6	50	監事	4.19	16
51	女子	4.16	31	52	職能	4.14	10
53	範囲	4.14	22	54	被用者年金	4.12	41
55	資金	4.09	6	56	障害福祉年金	4.07	67
57	中途	4.02	12	58	地域	4.00	10
59	夫	3.97	77	60	職務	3.94	19
61	氏名	3.90	6	62	医療	3.88	10

63	マツサージ指圧	3.85	3	64	きゆう	3.85	3
65	柔道	3.85	3	66	食品衛生	3.85	3
67	児童福祉	3.85	3	68	はり	3.85	3
69	技術	3.82	8	70	世帯	3.78	19
71	国	3.78	10	72	厚生大臣	3.76	10
73	限り	3.75	4	74	義務教育	3.73	10
75	犯罪	3.73	10	76	月数	3.72	174
77	全国	3.70	23	78	事務所	3.69	7
79	会社	3.69	7	80	学識	3.65	11
81	罰金	3.61	10	82	学生	3.58	25
83	福祉	3.57	5	84	子	3.48	157
85	相当	3.46	6	86	所得	3.42	57
87	任期	3.39	10	88	発起人	3.36	9
89	妻	3.33	88	90	共済水産	3.29	5
91	生命保険会社	3.29	5	92	精神保健	3.28	1
93	歯科技工	3.28	1	94	臨床	3.28	1
95	老人福祉	3.28	1	96	身体障害	3.28	1
97	狂犬病	3.28	1	98	母子保健	3.28	1
99	社会福祉事業	3.28	1	100	食	3.28	2
101	衛生	3.28	2	102	業法	3.28	1
103	環境	3.28	1	104	知的障害	3.28	1
105	精神障害	3.28	1	106	技師	3.28	2
107	患者	3.28	1	108	結核	3.28	1
109	と畜場	3.28	1	110	公衆浴場	3.28	1
111	旅館業法	3.28	1	112	期間	3.23	846
113	船員任意	3.23	8	114	負	3.23	5
115	各項	3.22	10	116	住所	3.17	38
117	法人	3.16	21	118	例	3.07	157
119	費用	3.04	63	120	つて生計	3.04	41
121	原因	3.04	4	122	地方厚生局長	3.04	5
123	年金保険	3.04	21	124	地区	3.03	7
125	国税	3.03	8	126	男子	3.01	13
127	責任	2.98	4	128	都道府県	2.96	5
129	地方自治	2.92	9	130	銀行	2.92	3
131	社会保険庁長官	2.91	14	132	平成	2.90	167
133	祖父	2.89	10	134	基準月	2.85	8
135	効率	2.85	6	136	母子状態	2.84	7
137	払	2.82	8	138	基礎	2.81	14
139	債権	2.80	11	140	財産	2.79	6
141	議事	2.78	6	142	別段	2.77	13
143	刑法	2.76	5	144	職員	2.75	33
145	国民年金手帳	2.75	6	146	天災	2.72	9
147	資料	2.68	3	148	状態	2.68	127

149	基礎年金番号	2.66	3	150	公社	2.65	4
151	生死	2.65	3	152	価	2.63	10
153	老齢厚生年金	2.62	14	154	収入	2.61	4
155	制度	2.59	5	156	生計	2.56	55
157	資格要件	2.56	9	158	基準	2.56	43
159	会員	2.56	9	160	年金事業	2.55	2
161	申立て	2.54	10	162	法令	2.52	62
163	所在地	2.52	5	164	社会保険事務所	2.52	4
165	名称	2.52	10	166	書面	2.51	5
167	総務	2.48	6	168	父	2.47	37
169	年数	2.46	2	170	事業主	2.46	8
171	欄	2.44	47	172	郵政民営	2.41	2
173	政府	2.41	64	174	行政法人福祉医療機構	2.40	6
175	故意	2.39	6	176	他	2.39	57
177	財務大臣	2.39	6	178	胎児	2.38	6
179	つた子	2.38	5	180	物件	2.38	5
181	虚偽	2.36	12	182	長	2.36	4
183	資産	2.35	2	184	遺児年金	2.35	37
185	前々	2.34	4	186	母子年金	2.34	62
187	明治	2.34	20	188	最初	2.33	13
189	従前	2.33	127	190	同種	2.32	4
191	支払	2.32	12	192	金銭	2.31	4
193	旨	2.30	4	194	地方社会保険事務	2.30	4
195	年のうち	2.30	4	196	官公	2.30	3
197	過料	2.30	6	198	疾病	2.30	21
199	口座振替	2.29	7	200	保険協会	2.29	1
201	何人	2.29	1	202	在り方	2.29	9
203	郵便振替	2.27	1	204	郵便為替	2.27	1
205	郵便振替預り	2.27	1	206	郵便	2.27	1
207	地方公務員共済組合	2.26	13	208	下	2.24	5
209	日本国内	2.23	25	210	知識	2.22	4
211	能力	2.22	2	212	実務	2.22	2
213	附則別表	2.22	10	214	財源	2.21	5
215	数	2.19	40	216	程度	2.17	115
217	定数	2.16	4	218	目途	2.15	3
219	基準傷病	2.14	5	220	方法	2.14	15
221	利益	2.12	6	222	日本	2.12	4
223	基礎年金	2.11	22	224	懲役	2.10	5
225	趣旨	2.10	2	226	母	2.10	27
227	障害等級	2.09	53	228	口座	2.09	6
229	所在	2.06	5	230	厚生	2.05	247
231	日本年金機構	2.03	5	232	老齢基礎年金	2.03	98
233	罪	2.02	4	234	情報処理	2.02	2

235	国際	2.02	2	236	字句	2.01	9
237	私立学校教職員共済制度	2.01	3	238	地方公共団体	1.99	16
239	年金積立	1.98	7	240	地方厚生支局	1.98	3
241	社会	1.97	5	242	共済年金	1.96	22
243	船舶	1.95	8	244	国籍	1.95	2
245	不適当	1.94	2	246	臨時	1.94	4
247	効果	1.94	4	248	上級行政	1.92	6
249	老齢	1.92	22	250	業者	1.91	6
251	納	1.91	26	252	行	1.91	7
253	番号	1.90	2	254	記号	1.90	2
255	残余財産	1.89	5	256	厚生年金保険事業	1.89	3
257	至	1.89	74	258	比率	1.87	18
259	保険組合	1.86	2	260	観点	1.86	6
261	章	1.85	7	262	金融商品	1.84	4
263	全額	1.83	20	264	行方	1.83	11
265	劇	1.83	2	266	毒物	1.83	2
267	水道	1.83	2	268	薬事	1.83	2
269	農林漁業団体職員共済組合	1.82	5	270	地方社会保険事務局長	1.82	3
271	私学教職員共済制度	1.82	7	272	引継ぎ	1.81	2
273	次号	1.81	14	274	入管	1.81	2
275	警察職員	1.81	4	276	表	1.81	76
277	学校	1.81	2	278	大学	1.81	2
279	貸付け	1.80	3	280	異	1.79	2
281	年	1.79	66	282	総務庁	1.79	9
283	共済組合	1.78	78	284	遺族厚生年金	1.78	6
285	農業	1.78	14	286	対象配偶	1.77	6
287	期限	1.77	13	288	市町村	1.76	22
289	農林共済組合	1.76	7	290	財政	1.75	25
291	大臣	1.74	165	292	直近	1.74	11
293	便益	1.74	1	294	者	1.74	1
295	本邦	1.74	3	296	都道府県知事	1.73	16
297	医師	1.73	4	298	歯科医師	1.73	4
299	被害金額	1.72	1	300	震災	1.72	1
301	火災	1.72	1	302	住宅	1.72	1
303	家財	1.72	1	304	風水害	1.72	1
305	価格	1.72	1	306	障害共済年金	1.71	12
307	老齢年金	1.71	110	308	地方議会議員共済	1.70	1
309	求め	1.70	5	310	過失	1.69	2
311	親族	1.69	16	312	初診	1.69	83
313	衛視	1.69	4	314	掛金	1.69	9
315	端数	1.68	90	316	積立	1.68	22
317	物価	1.67	6	318	船員保険	1.66	41
319	ロ	1.66	15	320	兄弟姉妹	1.65	2

321	祖父母	1.65	2	322	父母	1.65	2
323	所得税法	1.65	17	324	別表	1.64	30
325	総理府	1.63	5	326	債務	1.63	2
327	補欠	1.63	4	328	前任	1.63	4
329	残任期間	1.63	4	330	過半数	1.63	2
331	決し	1.63	2	332	可否同数	1.63	2
333	社会経済	1.62	1	334	被用者	1.62	2
335	企業	1.62	3	336	仕組み	1.62	1
337	公平	1.62	1	338	形態	1.62	2
339	年齢	1.62	3	340	割合	1.61	21
341	障害状態	1.61	3	342	政令	1.60	250
343	傷病	1.59	90	344	寡婦年金	1.59	20
345	条約	1.58	3	346	難民	1.58	4
347	至つた子	1.58	2	348	なつ	1.58	159
349	法律附則	1.58	57	350	納期	1.57	6
351	国民年金制度	1.56	3	352	地方	1.56	4
353	国民年金事業	1.56	25	354	新法附則	1.55	7
355	航空機	1.55	6	356	法附則別表	1.53	1
357	障害厚生年金	1.53	10	358	利便	1.52	3
359	有無	1.52	11	360	見直し	1.52	4
361	戸籍	1.52	3	362	年金制度	1.51	15
363	行政事務	1.51	2	364	行政法人	1.49	6
365	責め	1.49	3	366	任	1.49	3
367	上欄	1.48	24	368	私立学校教職員共済	1.46	11
369	社会保険	1.46	9	370	時効	1.45	15
371	実態	1.45	2	372	組合	1.45	57
373	法附則	1.45	43	374	金額	1.44	16
375	帳簿	1.44	6	376	適正	1.43	2
377	昭和	1.43	494	378	会	1.43	4
379	部分	1.43	59	380	通則	1.43	7
381	遺族	1.42	19	382	法制	1.42	2
383	日本国籍	1.41	4	384	書	1.40	63
385	総数	1.40	5	386	国庫	1.39	15
387	不服	1.39	8	388	後段	1.39	3
389	名	1.39	2	390	当事	1.38	3
391	児童	1.37	18	392	任務	1.37	2
393	現況	1.36	12	394	見通し	1.36	13
395	被告	1.36	2	396	各月	1.36	14
397	余裕	1.35	2	398	人格	1.35	4
399	社団	1.35	5	400	農林共済年金	1.35	3
401	共済	1.34	3	402	繰入れ	1.33	1
403	繰入金	1.33	1	404	財政投融資	1.33	2
405	国民年金基金	1.32	45	406	公共団体	1.31	1

407	原資	1.31	3	408	特例	1.31	7
409	出入国	1.31	3	410	国立大学法人	1.31	2
411	学校法人	1.31	1	412	公立大	1.31	1
413	私立学校	1.31	1	414	申	1.29	63
415	刑事	1.29	4	416	年金数理	1.28	5
417	水準	1.28	5	418	前段	1.28	1
419	旧法	1.27	7	420	事情	1.27	27
421	リ	1.27	1	422	外国	1.26	5
423	条本文	1.25	2	424	大正	1.25	10
425	おそれ	1.25	2	426	書中	1.24	4
427	遺族共済年金	1.23	7	428	特例事務法人	1.21	4
429	つて	1.21	84	430	事業年度	1.20	2
431	法人税	1.20	2	432	所得税	1.20	2
433	公債	1.20	3	434	保険	1.20	1290
435	順位	1.19	7	436	始	1.19	1
437	取消し	1.19	8	438	身分	1.18	6
439	価額	1.17	2	440	率	1.17	33
441	年度	1.17	41	442	年金基金	1.17	5
443	差額	1.17	3	444	出	1.16	5
445	厚生年金保険制度	1.16	2	446	農林漁業団体職員共済組合制度	1.16	2
447	障害年金	1.16	68	448	要件	1.16	62
449	総額	1.16	4	450	行つた障害基礎年金	1.15	2
451	税制	1.15	3	452	先	1.14	3
453	職権	1.14	1	454	障害	1.13	155
455	ぼつ	1.13	2	456	乙年金	1.13	1
457	甲年金	1.13	3	458	年金通則	1.12	5
459	過誤払	1.11	4	460	場所	1.11	2
461	日時	1.11	2	462	区域	1.11	2
463	暇	1.11	2	464	事由	1.11	86
465	人	1.11	3	466	本条	1.11	1
467	半数	1.11	2	468	つて年金	1.09	1
469	至つた孫	1.09	1	470	国税通則	1.09	1
471	実地	1.09	1	472	原子爆弾小頭	1.08	1
473	原子爆弾	1.08	3	474	保健	1.08	1
475	年金福祉事業団	1.08	1	476	予算	1.07	3
477	国民年金原簿	1.06	1	478	国家公務員共済組合	1.06	25
479	戦傷病者	1.05	2	480	支障	1.05	1
481	最低	1.05	1	482	段階	1.05	1
483	具体	1.05	1	484	少子化	1.05	1
485	配偶	1.03	56	486	後任	1.02	2
487	旧法障害年金	1.02	3	488	国民	1.02	7
489	残余	1.01	5	490	国家公務員	1.01	3
491	厚生年金保険	1.00	170	492	国税庁長官	1.00	2

493	中間法人	0.99	1	494	犯罪収益	0.99	1
495	法律別表	0.99	1	496	地位	0.99	3
497	民法	0.97	3	498	日本国民	0.97	8
499	新法	0.97	6	500	つた	0.96	15
501	障害年金附則	0.95	1	502	異議	0.95	2
503	不利益	0.95	2	504	訴え	0.95	2
505	全国市町村職員共済組合	0.94	1	506	最後	0.94	5
507	地方社会保険医療	0.94	2	508	会長	0.94	2
509	専門委員	0.94	2	510	社会保険医療	0.94	1
511	地方税財源	0.94	1	512	自主	0.94	1
513	経済情勢	0.94	1	514	方途	0.94	1
515	役割	0.94	1	516	つて厚生年金保険	0.94	2
517	次順位	0.93	1	518	先順位	0.93	1
519	障害基礎年金	0.93	129	520	期間附則	0.93	2
521	船員組合	0.93	5	522	財務	0.93	2
523	災害福祉年金	0.92	1	524	別	0.92	2
525	坑内	0.92	1	526	船員	0.92	2
527	鉱業	0.92	1	528	被用者年金保険	0.92	6
529	条項	0.91	1	530	名目手取り賃金	0.91	4
531	私立学校教職員共済組合	0.90	3	532	警察	0.90	2
533	効力	0.89	33	534	改任	0.89	1
535	地方公務員	0.89	20	536	末日	0.88	2
537	機会	0.88	1	538	陸軍共済組合	0.87	2
539	符号	0.87	1	540	生徒	0.87	3
541	無料	0.86	1	542	理念	0.85	1
543	つて国民	0.85	1	544	日本国憲法	0.85	1
545	特例基準	0.85	1	546	商業手形	0.85	1
547	性質	0.84	1	548	給料	0.84	1
549	証票	0.84	2	550	症状	0.84	2
551	附則	0.84	363	552	標準報酬月額	0.84	2
553	標準賞与	0.84	2	554	市場	0.83	1
555	既往	0.83	1	556	日数	0.83	1
557	つた傷病	0.82	1	558	月	0.82	281
559	標準報酬	0.82	4	560	経済社会	0.81	1
561	動向	0.81	1	562	我が国	0.81	1
563	期	0.81	1	564	国家行政	0.80	1
565	次項	0.80	26	566	医療保険制度	0.79	1
567	視点	0.79	1	568	体制	0.79	1
569	任意	0.77	3	570	厚生労働省	0.77	1
571	全力	0.77	1	572	公務	0.77	1
573	財政方式	0.77	1	574	地方税法	0.77	2
575	態様	0.76	1	576	生年月日	0.76	1
577	偽り	0.75	2	578	手段	0.75	2

579	災害	0.75	6	580	改	0.73	1
581	国会	0.73	1	582	制度全般	0.73	1
583	税	0.73	1	584	被疑	0.73	1
585	権利	0.73	78	586	不作為	0.72	1
587	恩給	0.71	8	588	実績	0.71	1
589	形	0.71	1	590	額	0.69	572
591	遺族年金	0.69	16	592	厚生保険	0.69	2
593	特例遺族年金	0.69	2	594	南西諸島官公	0.68	1
595	軍人	0.68	1	596	住民基本台帳	0.68	3
597	その子	0.68	1	598	地方分権	0.68	1
599	種類	0.67	1	600	秘密	0.67	2
601	行つた障害	0.67	1	602	条例	0.67	5
603	基準障害	0.67	2	604	遺族基礎年金	0.65	66
605	辞令	0.65	1	606	厚生省	0.65	1
607	保険福祉	0.65	2	608	年金福祉	0.65	1
609	行政法人年金	0.65	2	610	手数料	0.65	1
611	特例年金	0.63	1	612	農林年金	0.63	1
613	至つた	0.63	1	614	次	0.62	177
615	前項本文	0.62	1	616	間	0.62	89
617	消防職員	0.60	1	618	年金条例職員期間	0.60	1
619	議長	0.60	4	620	社会経済情勢	0.60	1
621	類型	0.60	1	622	あり方	0.60	1
623	文字	0.60	2	624	物価指数附則	0.59	2
625	財団	0.59	1	626	級	0.58	1
627	用語	0.57	1	628	意義	0.57	1
629	国民年金事務組合	0.57	2	630	別表住民基本台帳	0.57	1
631	年金	0.57	425	632	つた保険	0.54	1
633	対象	0.54	20	634	養子縁組	0.51	3
635	農林共済	0.51	3	636	積立て	0.51	1
637	監獄	0.51	2	638	様式	0.51	1
639	体系	0.51	1	640	責	0.50	1
641	相互	0.50	1	642	国民年金	0.50	681
643	国税局長	0.50	1	644	国税局	0.50	1
645	税務署	0.50	1	646	条各項	0.50	1
647	順序	0.49	2	648	職	0.49	1
649	額附則	0.48	1	650	財産所在地	0.47	1
651	地方公務員共済	0.47	1	652	収支	0.46	1
653	年月日	0.46	1	654	国家公務員災害	0.45	1
655	区長	0.43	3	656	特例基準割合	0.43	2
657	取扱い	0.42	1	658	便宜	0.42	1
659	抜本	0.42	2	660	市町村税	0.42	1
661	つて遺族基礎年金	0.42	2	662	年度末	0.41	1
663	厚生年金	0.41	1	664	執達吏規則	0.40	1

665	年度政令	0.39	1	666	寡婦	0.38	1
667	達	0.37	1	668	法人船員組合	0.37	2
669	電子情報処理	0.37	1	670	合理	0.36	1
671	通	0.35	1	672	原告	0.34	1
673	裁判所	0.34	1	674	公課	0.34	1
675	標準	0.34	1	676	租税	0.34	1
677	結了	0.34	1	678	至つた傷病	0.34	1
679	立入り	0.34	1	680	立入検査	0.34	1
681	行政事件	0.34	1	682	議院	0.33	1
683	前条	0.32	108	684	法	0.31	311
685	他方	0.30	1	686	障害基礎年金附則	0.30	1
687	至つた後こ	0.30	1	688	地方税	0.30	1
689	特権	0.30	1	690	うえ	0.30	1
691	細則	0.30	1	692	月間	0.29	1
693	前項	0.27	271	694	都市	0.26	1
695	並び	0.26	1	696	市	0.26	1
697	小数点	0.26	3	698	他人	0.26	2
699	つて世帯	0.25	2	700	地方事務	0.25	1
701	国会議員	0.25	1	702	項	0.25	244
703	基本方針	0.24	2	704	条	0.20	278
705	日本銀行	0.18	1	706	労役	0.17	1
707	理事長又	0.17	1	708	裁判	0.17	1
709	年額	0.17	1	710	節	0.17	1
711	対象月数	0.17	1	712	財団法人	0.15	1
713	目次	0.09	1	714	官報	0.09	1
715	少年院	0.09	1	716	保険料率	0.09	1

表 4.2: 労働基準法の χ^2 値

順位	ラベル	χ^2 値	頻度	順位	ラベル	χ^2 値	頻度
1	従前	8.26	45	2	附則	6.32	51
3	政令	6.21	41	4	罰則	4.68	25
5	機関	3.77	10	6	局長	3.15	17
7	事務	3.14	10	8	職員	3.12	12
9	法律	2.87	255	10	年数	2.87	8
11	上欄	2.87	3	12	欄	2.87	5
13	例	2.80	58	14	相当	2.75	6
15	農林水産省	2.72	2	16	外務省	2.72	2
17	法務省	2.72	2	18	都道府県	2.57	13
19	国	2.50	7	20	過半数	2.40	58
21	行政	2.35	14	22	表	2.28	11
23	字句	2.27	4	24	申立て	2.18	12
25	効力	2.18	12	26	組合	2.16	54
27	地方公共団体	2.15	9	28	賃金	2.07	65
29	辞令	1.85	2	30	公社	1.81	4
31	ガス	1.80	3	32	場所	1.78	6
33	内閣	1.76	3	34	会長	1.74	3
35	昭和	1.72	22	36	平成	1.69	34
37	府	1.66	4	38	事業	1.65	86
39	郵政民営	1.63	2	40	別表	1.62	10
41	厚生労働省	1.62	6	42	対象業務	1.61	10
43	地方自治	1.61	6	44	支払	1.60	3
45	女性	1.58	24	46	児童	1.57	14
47	銀行	1.54	1	48	郵便振替	1.54	1
49	郵便	1.54	1	50	郵便為替	1.54	1
51	郵便振替預り	1.54	1	52	政府	1.52	10
53	災害	1.51	18	54	環境省	1.46	1
55	総務	1.46	1	56	財務省	1.46	1
57	文部科学	1.46	1	58	経済産業	1.46	1
59	部局	1.46	3	60	国土交通省	1.46	1
61	書中	1.45	1	62	厚生	1.44	74
63	上級行政	1.44	6	64	理由	1.44	7
65	疾病	1.43	11	66	通常	1.42	10
67	大臣	1.39	13	68	妊産婦	1.38	8
69	条	1.36	63	70	新法	1.30	25
71	旧法	1.29	9	72	週	1.28	2
73	福祉	1.28	14	74	業務	1.28	43
75	状況	1.26	11	76	大蔵省	1.26	1
77	厚生省	1.26	1	78	通商産業省	1.26	1
79	総理府	1.26	1	80	自治省	1.26	1
81	文部省	1.26	1	82	日本工業標準	1.26	1
83	建設省	1.26	1	84	郵政省	1.26	1

85	運輸省	1.26	1	86	重量	1.26	2
87	機会	1.25	5	88	省令	1.25	63
89	原料	1.24	2	90	粉末	1.24	1
91	じんあい	1.24	1	92	毒劇薬	1.24	1
93	放射線	1.24	1	94	材料	1.24	2
95	高圧	1.24	1	96	高温	1.24	1
97	毒劇	1.24	1	98	所定	1.22	8
99	間	1.21	27	100	前項	1.19	56
101	衛生	1.17	7	102	書面	1.17	18
103	寄宿舎	1.15	13	104	条件	1.14	15
105	期日	1.14	7	106	中央	1.14	2
107	法	1.13	14	108	総額	1.13	6
109	坑内	1.13	5	110	名称	1.12	3
111	如何	1.12	3	112	社会	1.10	3
113	至	1.10	4	114	労働省	1.09	5
115	権利	1.09	6	116	つて	1.09	34
117	後見人	1.08	4	118	親権	1.08	4
119	事件	1.08	5	120	通貨	1.08	4
121	署長	1.08	3	122	日数	1.07	23
123	他	1.06	4	124	行政官庁	1.05	39
125	未成年	1.02	6	126	事業主	1.02	11
127	率	1.01	5	128	障害	1.01	4
129	異議	1.01	4	130	限度	0.99	12
131	身分	0.98	3	132	職業	0.97	9
133	信条	0.96	2	134	国籍	0.96	2
135	身体	0.95	2	136	方法	0.95	8
137	当たり	0.94	9	138	事由	0.94	17
139	内部部局	0.94	2	140	局	0.94	2
141	所掌	0.94	2	142	並び	0.94	2
143	公共団体	0.93	1	144	男女	0.93	4
145	分野	0.93	4	146	目的	0.93	5
147	行	0.92	3	148	委員	0.91	21
149	休暇	0.90	4	150	家族	0.90	6
151	分限	0.90	4	152	金額	0.89	11
153	同種	0.88	1	154	月	0.88	1
155	法令	0.86	6	156	保険	0.86	7
157	産前	0.86	2	158	下級	0.85	3
159	官庁	0.85	3	160	クレーン	0.84	1
161	取りはずし	0.84	1	162	動力	0.84	3
163	ベルト	0.84	1	164	機械	0.84	2
165	ロープ	0.84	1	166	取付け	0.84	1
167	単位	0.83	2	168	雇	0.82	4
169	政策	0.82	4	170	利便	0.81	1

171	視点	0.81	1	172	体制	0.81	1
173	社会保険	0.81	1	174	在り方	0.81	1
175	効率	0.81	1	176	年金制度	0.81	1
177	医療保険制度	0.81	1	178	各日	0.81	3
179	取扱い	0.81	3	180	数	0.81	3
181	次	0.79	30	182	事情	0.78	3
183	動向	0.78	5	184	子	0.78	2
185	範囲	0.76	34	186	観点	0.76	2
187	地方	0.76	1	188	地区職業	0.76	1
189	地方最低賃金	0.76	1	190	地方家内	0.76	1
191	地方職業	0.76	1	192	金品	0.74	2
193	不利益	0.74	3	194	女子	0.73	3
195	額	0.73	8	196	職務	0.72	4
197	比率	0.72	3	198	取扱	0.72	2
199	罰金	0.71	5	200	年少	0.71	3
201	臨時	0.69	10	202	技能	0.69	3
203	なつ	0.68	4	204	差	0.67	1
205	繁閑	0.67	1	206	寄宿舎規則	0.66	3
207	程度	0.66	1	208	官吏	0.64	3
209	権限	0.63	2	210	坑内業務	0.63	1
211	職業能力	0.63	1	212	指針	0.62	3
213	業種	0.62	2	214	規模	0.62	2
215	行政事務	0.62	1	216	利子	0.62	4
217	要員	0.61	1	218	実態	0.61	1
219	基礎	0.61	2	220	時効	0.60	3
221	責	0.60	4	222	請負人	0.60	6
223	賃金支払	0.59	2	224	教養	0.58	1
225	年報	0.58	1	226	項	0.58	46
227	天災事変	0.58	2	228	地方税財源	0.58	1
229	役割	0.58	1	230	経済情勢	0.58	1
231	方途	0.58	1	232	自主	0.58	1
233	秘密	0.57	2	234	行政事件	0.57	2
235	国家行政	0.57	1	236	雇入	0.57	3
237	下	0.56	2	238	問題	0.56	1
239	特性	0.56	1	240	賃金台帳	0.56	2
241	手段	0.56	2	242	別段	0.55	7
243	対象期間	0.55	12	244	利益	0.55	2
245	割増賃金	0.55	13	246	長	0.55	1
247	費用	0.54	3	248	文書	0.53	1
249	行つた	0.53	1	250	利率	0.52	3
251	金融機関	0.52	1	252	民法	0.52	2
253	制度	0.52	1	254	次項	0.51	6
255	作業場	0.50	2	256	公	0.50	2

257	左	0.50	4	258	賞与	0.50	2
259	見直し	0.50	1	260	地方分権	0.50	1
261	記号	0.49	1	262	家庭	0.49	1
263	不作為	0.49	1	264	民事	0.49	1
265	即時	0.49	2	266	事項	0.48	56
267	男性	0.48	2	268	有給休暇	0.48	21
269	管内	0.48	1	270	書	0.48	19
271	期間	0.47	90	272	市町村	0.47	2
273	要旨	0.47	1	274	種類	0.47	2
275	規則	0.46	15	276	名簿	0.46	2
277	前条	0.46	35	278	最初	0.46	9
279	任期	0.46	1	280	半数	0.46	1
281	書類	0.45	2	282	徒弟	0.44	1
283	見習	0.44	1	284	懲役	0.43	3
285	部分	0.43	37	286	積立	0.43	1
287	下請負人	0.43	2	288	手数料	0.43	1
289	旅費	0.43	3	290	戸籍	0.42	3
291	演劇	0.42	2	292	映画	0.42	2
293	商店	0.42	1	294	黄燐燐寸	0.42	1
295	工場	0.42	1	296	工業	0.42	1
297	最低年齢	0.42	1	298	外	0.41	2
299	給料	0.41	1	300	対	0.41	1
301	未然	0.39	1	302	数次	0.38	2
303	つて業務	0.37	1	304	過失	0.37	2
305	下級官庁	0.37	1	306	医師	0.37	2
307	支払期日	0.37	1	308	既往	0.37	1
309	生命	0.37	1	310	定員	0.37	1
311	防湿	0.37	1	312	採光	0.37	1
313	風紀	0.37	1	314	基準局	0.37	1
315	価額	0.37	1	316	公衆	0.37	1
317	不便	0.37	1	318	官公	0.36	1
319	国家公務員	0.36	1	320	地方公務員	0.36	1
321	公務	0.36	1	322	代理人	0.36	1
323	使用人	0.36	1	324	本条	0.36	1
325	学校	0.36	1	326	事故	0.35	1
327	標準報酬日額	0.35	1	328	危害	0.34	1
329	当事	0.34	1	330	最低	0.34	1
331	定め	0.33	28	332	法律附則	0.33	1
333	日日雇い	0.33	2	334	章	0.33	2
335	基準	0.33	133	336	通常所定	0.33	1
337	地位	0.32	1	338	旨	0.32	1
339	電気	0.31	1	340	他人	0.31	1
341	各種動力	0.31	1	342	水道	0.31	1

343	何人	0.31	1	344	業	0.31	1
345	次号	0.31	5	346	自治	0.31	1
347	室長	0.31	1	348	寮長	0.31	1
349	役員	0.31	1	350	精神	0.31	1
351	意思	0.31	1	352	苦情	0.31	2
353	専門	0.31	2	354	知識	0.31	2
355	至つた女子	0.31	1	356	未払金	0.31	1
357	裁判所	0.31	1	358	支払能力	0.30	1
359	資料	0.30	1	360	団体	0.30	1
361	直前	0.29	1	362	賃金締切	0.29	2
363	地域	0.28	1	364	生年月日	0.28	1
365	氏名	0.28	1	366	履歴	0.28	1
367	定期	0.27	1	368	きまつ	0.26	1
369	産業	0.26	1	370	つて打切	0.26	1
371	社会保険労務	0.26	2	372	罪	0.26	1
373	刑事	0.26	1	374	司法警察官	0.26	1
375	打切	0.25	1	376	戸籍事務	0.25	1
377	坑口	0.25	2	378	妨げ	0.25	1
379	掌	0.25	1	380	つた	0.25	1
381	公民	0.25	1	382	家事	0.25	1
383	無料	0.25	1	384	具体	0.25	1
385	中小事業主	0.25	1	386	議院	0.25	1
387	生後	0.25	1	388	生理	0.25	2
389	生	0.25	2	390	内容	0.25	1
391	不服	0.25	1	392	暇	0.25	1
393	事態	0.25	1	394	事後	0.25	1
395	責め	0.25	1	396	支障	0.25	1
397	事務所	0.25	1	398	債権	0.24	1
399	貸	0.24	1	400	遺族	0.24	2
401	葬祭	0.24	2	402	細目	0.24	1
403	健康上	0.23	1	404	議事	0.22	2
405	帳簿	0.21	1	406	民間事業	0.21	1
407	公聴	0.21	1	408	草案	0.21	1
409	信書	0.21	1	410	公益	0.21	1
411	船員	0.21	2	412	特例	0.21	1
413	証票	0.21	1	414	資格	0.20	1
415	裁判	0.19	1	416	児童附則	0.19	1
417	対等	0.19	1	418	立場	0.19	1
419	週所定	0.19	1	420	地方事務	0.19	1
421	満	0.19	1	422	職権	0.19	1
423	前項但書後段	0.19	1	424	事業場外	0.19	1
425	住居	0.18	1	426	出来高払	0.18	1
427	定額	0.18	1	428	全額	0.18	1

429	最低賃金	0.18	1	430	最低基準	0.18	1
431	親族	0.17	1	432	家事使用人	0.17	1
433	年齢	0.16	1	434	勅	0.16	1
435	私生活	0.16	1	436	合理	0.16	1
437	受入	0.13	1	438	行方	0.13	1
439	人たる	0.13	1	440	農業	0.13	1
441	法附則	0.13	1	442	各項	0.13	1
443	義務	0.12	1	444	但書	0.12	1
445	節	0.12	1	446	異	0.10	1
447	不履行	0.06	1	448	林業	0.06	1

表 4.3: 建築基準法の χ^2 値

順位	ラベル	χ^2 値	頻度	順位	ラベル	χ^2 値	頻度
1	資格	6.18	119	2	公益	5.70	28
3	工業	5.20	16	4	業務	5.18	162
5	長	5.12	37	6	役員	4.88	19
7	罰則	4.84	39	8	利便	4.61	16
9	機関	4.34	168	10	居室	4.21	24
11	理由	4.18	30	12	特例容積	3.99	20
13	延べ面積	3.95	30	14	法人	3.92	24
15	事務所	3.76	27	16	部分	3.62	218
17	構造方法	3.59	38	18	基本方針	3.56	6
19	容積	3.55	59	20	工業地域	3.53	13
21	壁	3.46	21	22	商業地域	3.46	8
23	都道府県知事	3.35	96	24	職員	3.29	40
25	従前	3.22	57	26	型式部材	3.19	41
27	帳簿	3.15	14	28	期限	3.12	21
29	床	3.08	14	30	技術	3.06	45
31	開口	3.05	11	32	最高限度	3.04	31
33	附則	3.04	54	34	数値	3.00	37
35	床面積	2.99	19	36	間	2.98	23
37	地域	2.98	37	38	書類	2.91	23
39	名称	2.90	10	40	住居地域	2.89	23
41	階段	2.86	12	42	地区	2.86	194
43	近隣商業地域	2.81	7	44	新法	2.78	28
45	柱	2.77	18	46	敷地面積	2.69	39
47	高層住居	2.68	14	48	行	2.67	20
49	線	2.66	9	50	主事	2.62	134
51	例	2.61	52	52	最低限度	2.61	22
53	別表	2.59	49	54	用途地区	2.56	5
55	欄	2.56	22	56	所在地	2.54	15
57	沿道地区	2.54	39	58	長等	2.50	12
59	外壁	2.48	19	60	境界	2.48	22
61	国	2.48	31	62	物件	2.44	7
63	住所	2.44	12	64	沿道	2.41	15
65	階数	2.39	9	66	委員	2.38	34
67	借地	2.31	17	68	危害	2.23	6
69	事務	2.21	77	70	対象区域	2.19	50
71	方法	2.19	25	72	天井	2.17	6
73	割合	2.17	13	74	前面道路	2.11	28
75	窓	2.10	6	76	旧法	2.10	31
77	性能	2.06	39	78	都市	2.05	132
79	風致	2.01	21	80	区内	1.99	8
81	敷地境界	1.99	4	82	病院	1.97	5
83	寄宿舎	1.97	5	84	ろ	1.97	4

85	事業	1.91	29	86	罰金	1.90	9
87	効力	1.89	38	88	学校	1.88	5
89	歴史	1.88	17	90	屋根	1.88	12
91	い	1.87	10	92	法令	1.87	7
93	建ぺい率	1.85	27	94	身体	1.85	4
95	用途地域	1.83	17	96	図書	1.82	12
97	相当	1.82	10	98	表	1.81	9
99	交通	1.80	30	100	手数料	1.79	7
101	周囲	1.75	11	102	規則	1.75	3
103	法律附則	1.74	4	104	軒	1.74	9
105	業務区域	1.73	12	106	地階	1.73	7
107	なつ	1.72	26	108	特例敷地	1.70	19
109	防火	1.68	79	110	門	1.67	5
111	土地	1.67	133	112	火災	1.67	17
113	倉庫	1.65	3	114	生命	1.63	5
115	政府	1.63	9	116	模様替	1.63	29
117	現場	1.62	8	118	災害	1.62	10
119	防火地域	1.60	42	120	内容	1.59	16
121	防火壁	1.58	12	122	幅員	1.58	20
123	空地	1.56	16	124	日影	1.53	4
125	冬至	1.53	3	126	海	1.52	3
127	市町村	1.51	108	128	塀	1.51	5
129	範囲	1.50	69	130	次項	1.49	22
131	乙種防火地区	1.48	1	132	甲種防火地区	1.48	1
133	防火区域	1.48	1	134	臨時防火	1.48	1
135	空地地区	1.48	2	136	美観地区	1.48	2
137	不作為	1.48	7	138	公務	1.48	4
139	刑法	1.48	4	140	商業	1.47	4
141	中心線	1.46	4	142	権利	1.44	5
143	法	1.44	33	144	水平面	1.44	3
145	市街地	1.42	31	146	面積	1.41	31
147	申立て	1.39	10	148	郵政民営	1.39	2
149	道	1.38	16	150	耐火構造	1.38	6
151	地盤	1.37	17	152	国土交通省令	1.37	149
153	道路	1.37	54	154	条件	1.37	20
155	廊下	1.36	4	156	衛生	1.36	59
157	材料	1.36	19	158	風土	1.36	3
159	気候	1.36	3	160	品質	1.35	7
161	地方自治	1.35	9	162	景観地区	1.35	13
163	型式	1.34	18	164	都道府県	1.32	41
165	国土交通大臣	1.31	190	166	銀行	1.31	1
167	郵便振替預り	1.31	1	168	郵便振替	1.31	1
169	郵便	1.31	1	170	郵便為替	1.31	1

171	公社	1.31	2	172	川	1.30	3
173	各項	1.29	14	174	特例対象	1.29	4
175	大都市地域	1.28	6	176	定め	1.27	5
177	室	1.27	3	178	歩廊	1.27	4
179	採光面積	1.27	4	180	額	1.27	5
181	工場	1.26	2	182	権限	1.25	13
183	太陽	1.25	1	184	距離	1.24	24
185	工程	1.24	35	186	煙突	1.24	5
187	職	1.23	2	188	当	1.21	4
189	隣地境界	1.20	5	190	請負人	1.20	4
191	専門	1.19	4	192	実費	1.18	4
193	木造	1.18	4	194	引継ぎ	1.17	3
195	通常	1.17	9	196	所定	1.16	3
197	社会資本	1.16	2	198	支障	1.16	62
199	こんろ	1.16	2	200	浴室	1.16	2
201	かまど	1.16	2	202	器具	1.16	2
203	集落地区	1.15	14	204	事情	1.15	4
205	費用	1.15	8	206	高低	1.15	4
207	長又	1.14	5	208	天災	1.13	3
209	汚物	1.12	6	210	利害	1.12	8
211	識見	1.11	4	212	下	1.09	9
213	業法	1.08	5	214	代理人	1.08	4
215	消防署	1.08	5	216	地方	1.08	4
217	最小	1.07	2	218	別段	1.07	3
219	保安	1.07	6	220	氏名	1.07	3
221	耐火	1.06	21	222	便所	1.06	5
223	敷地	1.05	172	224	石綿	1.05	6
225	メートル	1.05	4	226	平成	1.04	12
227	定期	1.04	4	228	防火構造	1.04	4
229	壁面	1.03	42	230	規模	1.02	57
231	がけ	1.02	2	232	住居	1.01	15
233	種類	1.01	3	234	耐火性能	1.01	3
235	都道府県都市	1.00	8	236	事由	1.00	8
237	場所	0.99	4	238	基礎	0.99	8
239	書	0.99	15	240	条例	0.99	74
241	線路敷地	0.99	2	242	浄化槽	0.98	5
243	過半	0.98	4	244	住宅	0.98	37
245	ロ	0.97	24	246	内外	0.97	9
247	地下	0.97	4	248	不燃材	0.96	4
249	宅地建物	0.96	5	250	出入口	0.95	2
251	煙	0.95	2	252	スプリンクラー	0.95	2
253	消火栓	0.95	2	254	公共	0.95	5
255	木材	0.94	4	256	外国	0.94	9

257	章	0.94	14	258	防湿方法	0.94	2
259	市長	0.92	1	260	引継	0.92	1
261	漁業	0.92	1	262	農業委員	0.92	1
263	他	0.92	9	264	懲役	0.91	3
265	収入	0.91	2	266	団地	0.91	9
267	隣地	0.91	2	268	法律	0.91	332
269	行つた	0.90	4	270	と畜場	0.90	2
271	市場	0.90	2	272	親会社	0.89	3
273	期間	0.89	60	274	旧法別表	0.89	1
275	法別表	0.89	1	276	構え	0.89	6
277	秘密	0.88	4	278	基準法令	0.87	13
279	用途	0.87	49	280	意思	0.87	5
281	上級行政	0.86	6	282	前提	0.86	2
283	見地	0.86	2	284	採光	0.86	4
285	物質	0.85	4	286	目的	0.84	18
287	火の粉	0.82	2	288	消防	0.82	6
289	限度	0.82	45	290	本文	0.82	6
291	高架	0.81	3	292	議	0.81	5
293	ダンスホール	0.81	1	294	公衆浴場	0.81	1
295	劇場	0.81	1	296	百貨店	0.81	1
297	体育館	0.81	1	298	旅館	0.81	1
299	遊技	0.81	1	300	自動車車庫	0.80	2
301	至	0.79	11	302	病室	0.79	2
303	教室	0.79	2	304	用	0.79	3
305	室内	0.79	1	306	仕上げ	0.79	1
307	公園	0.78	2	308	広場	0.78	2
309	特性	0.78	2	310	ガス	0.78	2
311	後段	0.78	6	312	当事	0.78	3
313	局部	0.77	1	314	廻り舞台	0.77	1
315	間柱	0.77	1	316	間仕切壁	0.77	1
317	ひさし	0.77	1	318	屋外階段	0.77	1
319	小	0.77	1	320	揚げ床	0.77	1
321	おそれ	0.77	40	322	取消し	0.76	15
323	事項	0.76	82	324	書面	0.76	4
325	新規	0.76	6	326	環境	0.76	26
327	合理	0.76	10	328	応急	0.76	4
329	防火性能	0.75	2	330	持分	0.75	1
331	外国型式部材	0.74	6	332	地方公共団体	0.74	24
333	昭和	0.74	9	334	虚偽	0.74	7
335	通路	0.73	1	336	業者	0.73	3
337	市	0.73	4	338	建設省	0.69	2
339	基準容積	0.69	3	340	側	0.67	2
341	様式	0.67	2	342	可燃材料	0.67	3

343	プラスチック	0.67	3	344	並び	0.66	6
345	れんが	0.66	2	346	景観	0.66	8
347	屎尿浄化槽	0.66	3	348	地域歴史	0.66	4
349	間口	0.66	4	350	都市環境	0.66	1
351	防湿	0.66	2	352	出	0.66	5
353	機会	0.65	2	354	物品	0.64	1
355	可燃	0.64	1	356	公共団体	0.64	1
357	要件	0.63	4	358	店舗	0.63	2
359	求め	0.63	3	360	人	0.62	3
361	本条	0.62	1	362	使用人	0.62	1
363	火炎	0.62	2	364	炎性能	0.62	2
365	界壁	0.61	2	366	議事	0.61	2
367	商店	0.61	1	368	地震	0.61	2
369	自己	0.60	3	370	期日	0.60	6
371	火	0.60	3	372	情報	0.59	2
373	状態	0.59	3	374	鉄筋コンクリート	0.59	2
375	形態	0.58	2	376	寝室	0.58	1
377	体制	0.58	1	378	年金制度	0.58	1
379	在り方	0.58	1	380	効率	0.58	1
381	視点	0.58	1	382	保険	0.58	1
383	医療保険制度	0.58	1	384	社会保険	0.58	1
385	下水道	0.56	3	386	私道	0.56	3
387	ごみ	0.56	2	388	手段	0.55	9
389	福祉	0.54	2	390	長屋	0.54	2
391	任期	0.54	3	392	外	0.54	2
393	差	0.53	1	394	知識	0.53	3
395	規格	0.53	2	396	区域	0.53	252
397	市町村都市	0.52	3	398	広告塔	0.52	2
399	資料	0.52	2	400	近隣	0.51	1
401	住民	0.51	1	402	日用品	0.51	1
403	問題	0.50	2	404	街区内	0.50	3
405	刑	0.50	3	406	故意	0.50	3
407	公共下水道	0.50	2	408	身分	0.50	2
409	外側	0.49	2	410	証	0.49	3
411	役割	0.49	1	412	地方税財源	0.49	1
413	自主	0.49	1	414	経済情勢	0.49	1
415	方途	0.49	1	416	行政事務	0.49	1
417	大臣	0.47	1	418	電気	0.47	3
419	面	0.47	1	420	異議	0.46	2
421	台帳	0.46	4	422	相互	0.46	1
423	道路中心線	0.46	1	424	町村	0.46	1
425	別	0.46	2	426	訴え	0.45	2
427	塗	0.45	1	428	鉄	0.45	1

429	モルタル塗	0.45	1	430	不利益	0.45	2
431	会長	0.45	4	432	地位	0.45	1
433	官公	0.45	1	434	下小屋	0.45	1
435	停車場	0.45	1	436	材料置場	0.45	1
437	低層住宅	0.45	1	438	財産	0.44	2
439	条	0.44	132	440	保健所	0.44	2
441	省令	0.44	1	442	本邦	0.43	2
443	写し	0.43	1	444	耐震	0.42	1
445	建替え	0.42	1	446	各戸	0.42	1
447	小屋	0.42	1	448	指針	0.42	4
449	伝統	0.41	1	450	現状	0.41	1
451	文化財	0.41	1	452	証拠	0.41	2
453	項	0.41	188	454	つて構造	0.41	1
455	禁治産者	0.41	1	456	民法	0.41	1
457	保佐	0.41	1	458	職務	0.40	2
459	前項	0.40	208	460	構造	0.40	243
461	立入検査	0.40	2	462	終末	0.40	1
463	震災	0.39	1	464	風災	0.39	1
465	水災	0.39	1	466	つて資格	0.39	1
467	前条	0.39	83	468	川又	0.38	1
469	各階	0.38	1	470	方向	0.38	1
471	地上部分	0.38	1	472	地上	0.38	1
473	物置	0.37	1	474	茶室	0.37	1
475	納屋	0.37	1	476	あずま	0.37	1
477	経済	0.37	1	478	公衆衛生	0.37	1
479	写	0.37	1	480	コンクリート	0.37	1
481	鋼材	0.37	1	482	水面	0.37	1
483	被告	0.37	1	484	技術水準	0.37	1
485	国民	0.36	1	486	最低	0.36	1
487	行つた型式	0.36	1	488	報酬	0.36	1
489	街路	0.35	1	490	議会	0.35	1
491	盛土	0.35	1	492	塗壁	0.35	1
493	地面	0.34	1	494	汚水	0.34	2
495	前段	0.34	3	496	義務	0.34	2
497	都知事	0.34	1	498	都	0.34	1
499	農林物資	0.33	1	500	行政法人農林水産	0.33	1
501	技術センター	0.33	1	502	学識	0.33	2
503	臨時	0.33	2	504	石造	0.33	1
505	筋コンクリート	0.33	1	506	コンクリートブロック	0.33	1
507	地方分権	0.32	1	508	見直し	0.32	1
509	観点	0.32	1	510	客席	0.32	1
511	患者	0.32	1	512	上家	0.31	1
513	軌道	0.31	1	514	鉄道	0.31	1

515	跨線橋	0.31	1	516	プラットホーム	0.31	1
517	途上	0.31	1	518	湿度	0.31	1
519	応力	0.31	2	520	次	0.31	119
521	生年月日	0.31	1	522	方式	0.31	1
523	標識	0.31	3	524	日常	0.31	1
525	住戸	0.31	1	526	音	0.31	1
527	公衆	0.30	1	528	最後	0.30	1
529	最初	0.30	1	530	ウォーター	0.30	1
531	高架水槽	0.30	1	532	つた	0.30	1
533	収支	0.29	1	534	収支予算	0.29	1
535	権利利益	0.29	1	536	積雪	0.29	1
537	水圧	0.29	1	538	衝撃	0.29	1
539	風圧	0.29	1	540	適格	0.29	3
541	状況	0.29	35	542	スポーツ	0.29	1
543	帆布	0.29	1	544	親族	0.29	1
545	事件	0.29	1	546	臨時物資需給	0.29	1
547	宅地	0.28	1	548	意匠	0.28	1
549	津波	0.28	1	550	高潮	0.28	1
551	通風	0.28	2	552	程度	0.28	1
553	人口	0.28	1	554	国土交通省	0.28	1
555	空隙	0.28	1	556	借主	0.27	2
557	避雷針	0.27	1	558	冷房	0.27	1
559	効果	0.27	1	560	北海道開発局長	0.27	1
561	局長	0.27	1	562	不燃性能	0.26	1
563	準則	0.26	2	564	金額	0.26	1
565	境	0.25	1	566	屋内	0.25	1
567	屋上	0.25	1	568	看板	0.25	1
569	情状	0.25	1	570	取扱	0.25	1
571	周壁	0.24	1	572	鉄骨鉄筋コンクリート	0.24	1
573	鉄骨	0.24	1	574	但書	0.24	1
575	事務組合	0.24	1	576	役場事務組合	0.24	1
577	旨	0.24	2	578	工法	0.24	1
579	外力	0.23	1	580	力	0.23	1
581	工	0.23	1	582	行政機関	0.23	1
583	政令	0.23	226	584	見込み	0.23	1
585	行政	0.23	208	586	次号	0.22	3
587	補欠	0.22	1	588	前任	0.22	1
589	残任期間	0.22	1	590	後任	0.22	1
591	犯罪	0.22	2	592	数	0.22	2
593	時価	0.22	1	594	つて通常	0.22	1
595	口頭	0.21	1	596	雨水	0.21	1
597	下水	0.21	1	598	下水溝	0.21	1
599	職権	0.21	1	600	都市緑地	0.21	1

601	行つた性能	0.21	1	602	実務	0.21	1
603	幹線道路	0.21	1	604	所以	0.20	1
605	知	0.19	1	606	何人	0.19	1
607	つて	0.19	120	608	基準	0.18	316
609	原告	0.18	1	610	裁判所	0.18	1
611	事故	0.18	1	612	娯楽	0.18	1
613	被害	0.18	1	614	がけ崩れ	0.18	1
615	消防本部	0.17	1	616	ち	0.17	1
617	行政事件	0.17	1	618	図面	0.17	1
619	仕様	0.17	1	620	責任	0.17	1
621	議院	0.17	1	622	高齢	0.16	1
623	身体障害	0.16	1	624	意義	0.16	1
625	用語	0.16	1	626	特例	0.15	1
627	つて仮	0.15	2	628	過半数	0.14	1
629	会務	0.13	1	630	総理	0.13	1
631	証人	0.13	1	632	集落地域	0.13	1
633	最大	0.13	1	634	路面	0.13	1
635	商法	0.12	1	636	数量	0.12	1
637	用途相互	0.12	1	638	過失	0.12	1
639	市街	0.12	1	640	各種学校	0.09	1
641	節	0.09	1	642	過料	0.09	1
643	農業	0.08	1	644	法附則	0.08	1
645	財団法人	0.08	1	646	左	0.08	1
647	実質	0.08	1	648	住宅金融公庫	0.04	1
649	高架道路	0.04	1	650	障子	0.04	1
651	本人	0.04	1	652	寸	0.04	1